International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-022-01681-w

ORIGINAL ARTICLE

A descriptive human visual cognitive strategy using graph neural network for facial expression recognition

Shuai Liu[1,2,3] · Shichen Huang[3] · Weina Fu[3] · Jerry Chun-Wei Lin[4]

Received: 19 June 2022 / Accepted: 29 September 2022
© The Author(s) 2022

Abstract

In the period of rapid development on the new information technologies, computer vision has become the most common application of artificial intelligence, which is represented by deep learning in the current society. As the most direct and effective application of computer vision, facial expression recognition (FER) has become a hot topic and used in many studies and domains. However, the existing FER methods focus on deep learning to generate increasingly complex attention structures, so they are unable to consider the connotative relationship between different parts of facial expressions. Moreover, the human expression recognition method based on complex deep learning network has serious interpretability issues. Therefore, in this paper, a novel Graph Neural Network (GNN) model is proposed to consider the systematic process of FER in human visual perception. Firstly, a region division mechanism is proposed, which divides the face region into six parts to unify the selection of key facial features. On this basis, in order to better consider the connotative relationship between different parts of facial expression, a human visual cognition strategy is proposed, which uses the divided six regions to learn facial expression features, and evenly selects the key features with high reliability as graph nodes. In combination with the human regional cooperative recognition process, the connotative relationship (such as relative position and similar structure) between graph nodes is extracted, so as to construct the GNN model. Finally, the effect of FER is obtained by the modeled GNN model. The experimental results compared with other related algorithms show that the model not only has stronger characterization and generalization ability, but also has better robustness compared with state-of-the-art methods.

Keywords Facial expression recognition · Graph neural network · Local visual cognition · Regional collaborative recognition

1 Introduction

As one of the most important branches of artificial intelligence, computer vision has been widely used in recent years in many fields, such as safety inspection of industrial products, visual image positioning, medical imaging, object measurement, and 3D reconstruction [1–3]. In these fields, computer vision uses various imaging systems as sensitive input devices instead of visual organs. Computers are used to replace the brain in information processing and interpretation, and perceptual devices are used to perform perceptual recognition, perceptual tracking, and perceptual measurement of targets instead of biological eyes, which is increasingly becoming the main subject of graphics processing. Therefore, the ultimate goal of computer vision is to investigate how to enable machines to observe and understand the information of the external environment like biological or human visual systems [4].

✉ Weina Fu
fuwn@hunnu.edu.cn

✉ Jerry Chun-Wei Lin
jerrylin@ieee.org

[1] School of Educational Science, Hunan Normal University, Changsha, China

[2] Key Laboratory of Big Data Research and Application for Basic Education, Changsha, China

[3] College of Information Science and Engineering, Hunan Normal University, Changsha, China

[4] Department of Computer Science, Electrical Engineering, and Mathematical Science, Western Norway University of Applied Sciences, Bergen, Norway

Published online: 25 October 2022

FER as an important part of computer vision, develops with the development of facial recognition technology. FER is the most direct and effective emotion recognition mode. It has many applications in human–computer interaction, such as in safe driving, where it can predict the driver's driving condition based on the driver's facial expression to avoid accidents; in distance learning, the method and content of teaching can be adjusted based on students' facial expressions; in game production, the player's sense of immersion can be enhanced or reduced according to human emotions [5–8]. Facial recognition has a wide range of applications, so it has attracted more and more researchers to explore its various aspects in depth.

The generation of facial expressions is a very complicated process. If the human mentality and the external environment are not taken into consideration, what is presented to the observer is pure facial muscle movement, as well as changes in the shape and texture of the face. The static image shows the expression state of a single image when the expression occurs, and the dynamic image shows the expression movement process between multiple images. Thus, a distinction is made between the state and the processing object when the expression occurs. The FER algorithm is completed by four steps: image acquisition of the face region, face region detection, facial expression feature information extraction, and facial expression feature classification.

However, most of the FER methods focus on using deep learning to build increasingly complex attention structures to extract facial features or classify facial expressions, but cannot fully explore the connotative relationship between the relevant position in the face and the expression, resulting in no good generalization ability. At the same time, the FER method based on the depth feature cannot be interpreted well. To solve these two important problems, this paper proposes a new GNN model based on the FER system of human cognitive vision. The model extracts the keypoints in FER, establishes the associated information between keypoints through the graph structure, and integrates them into FER. Compared with the existing algorithms, the model not only has stronger characterization and generalization ability, but also has better robustness. The main original contributions of this paper can be summarized as follows:

1. A novel region partition mechanism is proposed to support GNN node selection. In this mechanism, the facial expression is partitioned into six local regions (top-left, top-right, top-center-left, top-center-right, center, and bottom) to achieve a uniform selection of key features of the face, and human visual cognition is incorporated into the FER method. Compared with the current FER methods, the auxiliary mechanism is more related to human visual cognition, and the GNN has a more accurate FER effect.
2. A human visual cognitive strategy is proposed to build a GNN model. This strategy realizes the two key processes of "local visual cognition" and "regional collaborative recognition". First, based on the region division mechanism, facial expression features are learned in 6 different regions, and key features are uniformly selected as graph nodes. Then, in combination with the human collaborative process, the hidden relationships between graph nodes (such as relative position, similarity structure, etc.) are extracted as the edge of the graph node to model the GNN structure. With this strategy, the GNN has stronger representation and generalization ability.
3. The GNN model is first applied to FER to describe visual cognitive strategies. The GNN model constructed by human visual cognition is suitable for describing the face information from the spatial dimension. By analyzing and modeling the relationship between different regions of the face, the width and depth of the network are increased, and the interpretability of the model is also ensured.
4. The new method introduced in this work is applied to the novel application domain. Using this method not only guarantees the effect of FER, but also broadens the application fields of neural network models. At the same time, it promotes the development of interpretability of cognitive science on GNN. In addition, this algorithm is more robust compared with other related algorithms.

The structure of this paper is as follows: Sect. 2 reviews the work related to human visual cognition, GNN, and FER; Sect. 3 focuses on the process of using the GNN based on cognitive vision strategies for FER; Sect. 4 makes a quantitative and qualitative analysis of the GNN model proposed in this paper; and Sect. 5 summarizes the work content of this paper and the outlook for future work.

2 Related works

2.1 Human visual cognition

Humans perceive the world through his organs of perception. Under normal circumstances, people form the cognition of external environment information through the five organs of perception: sight, hearing, smell, touch, and taste. However, among the five organs of human cognition, vision is the main channel through which humans detect and receive information. Research shows that of the five senses, about 70% of information is detected through vision. Thus, the visual cognitive system is the most commonly used sensory organ in humans to contact external information. It is capable of rationally analysing, associating, interpreting, and comprehending external phenomena and distinguishing the contours of information in the visual process [9, 10].

Vision is not only a psychological and physical perception, but also the source of creativity. Experience comes from understanding and analyzing the environment. And all human behavior begins with cognition, so without cognition, there is nothing. In visual cognition, human beings have certain characteristics. In general, the visual image is first perceived as a unified whole and then in terms of parts. That is, humans first "see" the whole of a composition and then the parts of the whole composition. Therefore, vision is the most important way for humans to perceive the objective world.

Given the increasingly complex evolution of information visualization and understanding, Green et al. [11] proposed a design guide for visual interface development by discussing the human cognitive modeling framework to optimize human–computer interaction. Based on the recent biological findings, Hong et al. [12] proposed a framework to imitate the active and dynamic learning and recognition process of primate visual cortex. Scott et al. [13] proposed a general cognitive architecture that closely combines symbol, spatial, and visual representations to realize the free movement of the cognition between feature patterns. James et al. [14] investigated the spatial, dynamic, and molecular properties of neural activity in a range of within-system cognitive tasks and found that neural activity converges to a low-dimensional manifold. This research furthered the understanding of cognitive brain tissue. Hong et al. [15] improved the theoretical framework of visual cognition by imitating the finer structure and function of primate visual cortex to realize position and scale recognition. Hideho et al. [16] investigated the relationship between difficulties in visual cognition and the eye movements of the target, and proposed a gaze movement system to estimate human cognition, which is widely used in the cognitive process of motor targets.

Our method realizes two key processes of "local visual cognition" and "regional cooperative recognition" by using human visual cognitive strategies. The face is divided into six parts to learn the features. Finally, the six parts are considered as a whole to judge the facial expression. It simulates the human learning process from part to whole.

2.2 Graph neural network

In computer science, the graph is a data structure composed of two parts: vertices and edges. The graph G can be described by the vertex set V and the contained edges E.

$$G = (V, E) \tag{1}$$

V is a non-empty finite set of vertices, and E is the set of edges between any two points in the vertex set V.

In modern terms, the representation of a graph generally uses circles to represent vertices, and line segments to represent edges, and one edge connects two different vertices. If there are no directional edges, edges are called undirected edges, and the graph is called an undirected graph, as shown in Fig. 1a. If some edges have directions, these edges are called directed edges, and the graph is called a directed graph, as shown in Fig. 1b.

The size of the disordered nodes in the graph is variable, and each node has a different number of neighbour nodes. Therefore, GNN is used in social networks, knowledge graphs, recommender systems, and even in life sciences [17–19]. For instance, Gori et al. [20] first proposed the concept of GNN and used it to process graph data structures. Dalibor et al. [21] extended the graph matching algorithm with a new method combining fuzzy logic and recurrent neural networks. This method was more adaptable to input noise than conventional neural networks based on nonfuzzy supervised learning. Shen et al. [22] proposed a novel similarity guided graph neural network to overcome the relationship information between different probe libraries. Marco et al. [23] combined the coding module of the recurrent neural network and the radial basis function network to estimate the unknown probability density function from unsupervised samples for graph classification and graph clustering, based on the graph structured data. Rusek et al. [24] proposed a novel GNN based network model that can understand the complex relationship between topology, routing and input traffic to accurately estimate the delay distribution and loss of each data packet in each source/target.

Our method completes facial expression recognition through GNN modelling of the face, and considers the relationship between different regions of the face through undirected graph, which improves the representation and generalization ability of the model and is interpretable.

2.3 Facial expression recognition

Facial expressions are the result of one or more actions or states of the facial muscles. These states of movement fully express the person's emotions to the observer. Facial expressions are a form of nonverbal communication. It is the primary means of expressing social information between humans, and it also occurs in most other mammals and some

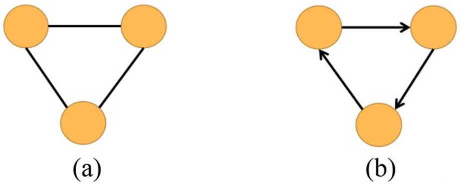

Fig. 1 Schematic diagram of the graph

other animal species. However, it is a difficult task to automatically recognize facial expressions from facial images [25–28].

The traditional FER first uses mathematical methods [29] (FastICA method, Gabor wavelet extraction method and so on) to extract facial expression features, such as keypoint information of facial features, facial texture features, facial histograms and other features. Then, classifiers (Naive Bayes classifier, Support Vector Machines, etc.) are used to classify the extracted information features and complete FER. For example, Ira et al. [30] used a naive Bayes classifier in the traditional FER to convert the Gaussian distribution of facial expression features into a Cauchy distribution, and used Gaussian to improve the naive Bayes classifier, which performs FER by learning the dependencies between the different facial expression features. Shan et al. [31] proposed a novel low-computation discriminative feature space based on a simple local binary pattern to represent the expression state of a face image. With this method, the features could not only be extracted faster, but also the face information could be stored effectively in a low-dimensional space. With the perfection of deep learning theory in recent years, deep learning technology has shown strong extraction ability in processing European-style image data, which has also promoted the rapid development of face recognition.

Unlike the traditional methods of FER, deep learning methods have a strong ability to extract features from images, especially the Convolutional Neural Network (CNN). The CNN is a convolutional kernel obtained by learning the feature types of facial expressions on the image, which naturally takes precedence over the features and classifiers of traditional methods. In these deep learning FER methods, multiple convolutional layers are first used to extract a large amount of key information about the facial expression features, and then a fully concatenated layer is used to perform nonlinear classification to realize FER. For example, Ali et al. [32] proposed a deep neural network architecture to address the shortcomings of inadequate representation of facial expressions by traditional features. Zeng et al. [33] developed a deep sparse autoencoder, introduced high-dimensional features, and proposed a novel FER framework to automatically distinguish facial expressions with high precision.

In the study of facial feature classification, Kamil et al. [34] proposed an entropy-based feature selection method applied to the 3D distance of facial features to realize FER by using the 3D distance between 83 points of the face as facial features. Khan et al. [35] proposed a new measurement for copolymer feature selection based on singular value decomposition to find the most salient areas in facial expressions to achieve accurate recognition. For FER, most methods use the entire image as the extraction region, ignoring the extraction of key areas of facial expressions. To solve this problem, He et al. [36] proposed a facial expression recognition method based on local binary patterns to fuse the key facial expression features. This method divides the facial expression into a part of the eyes, a part of the eyebrows, a part of the nose, and a part of the mouth. Then the feature information of these parts is obtained and combined to form a new feature. Then support vector machine and NN are used to classify the combined information for facial expression recognition. Shojaeilangari et al. [37] proposed an extreme sparse learning method that combines the discrimination ability of an extreme learning machine with the reconstruction of a sparse representation, which can achieve accurate FER in noisy signal areas and natural environments.

To solve the difficult feature extraction problem, Li et al. [38] proposed a unified probability concept based on dynamic Bayesian networks to represent changes in facial expression at the same level to detect the state of facial expression. Mohan et al. [39] designed a new type of deep CNN to study the geometric features of human faces, such as edges, curves, and straight lines. Then, a fractional fusion technology was used for facial expression recognition. Lee et al. [40] proposed a multimodal recurrent attention network integrating multimodal face information. This network could not only learn spatiotemporal attention information to add feature sets, but also use depth and popularity sequences as a priori advice for color sequence. And the emotion regions are selected to receive attention to achieve accurate recognition. To solve the problem caused by personal attributes and achieve better FER performance, Meng et al. [41] proposed a novel identity recognition CNN that can not only ensure that the features learned by the network remain unchanged under expression changes, but also that the identities remain unchanged. Andre et al. [42] proposed a combination of CNN and specific image preprocessing steps to solve the problem of FER. The network used the preprocessing technology to extract a large amount of feature data, and then used the deep structure to explore the presentation order of the facial expression feature data for FER. Sun et al. [43] proposed a multi-channel deep neural network to extract the optical flow from the changes between facial images with high expression and neutral facial images as the time information of facial expression. Then it used the horizontal images of gray emotional faces as spatial information to achieve FER.

3 Proposed method

In this section, starting from the insufficiency of existing FER algorithms based on deep learning, a novel GNN model is proposed to analyze the two key processes of human expression recognition in human visual cognition: “local visual recognition” and “regional collaborative recognition". First, this method selects cognitive key points based on human local visual cognition as graph nodes. Then, combined with the edge weights between the cognitive keypoints extracted by the regional collaboration process, they are used as the edge of the image. Finally, the graph nodes and graph edges extracted from the two processes of "local visual cognition" and "regional collaborative recognition" are connected to form a GNN model for FER.

The process of building a GNN model based on visual cognition for FER is shown in Fig. 2. Input a face image, divide the detected face into 6 regions (eyes, eyebrows, nose and mouth) according to the region division mechanism, select the more representative feature points of each region as the key feature points through the local visual recognition process, and finally connect the selected key nodes through the regional cooperative recognition process to build the GNN model, and improve the overall accuracy of the model through continuous training, Let the model accurately recognize facial expressions.

3.1 Insufficiency of existing deep facial expression recognition methods

The facial expressions of the human face are generally divided into expressions such as anger, disgust, fear, happiness, sadness, surprise, and natural. However, the change of each expression is reflected in multiple areas of the human face, such as eyebrows, eyes, nose, and mouth. Different facial expressions have different forms of expression in each area of the face. Therefore, the existing deep learning FER methods first perform preprocessing operations after inputting a face image or video. In the preprocessing process, the depth detection algorithm is used to divide the facial features Q into eyebrows Q_1, eyes Q_2, nose Q_3, and mouth Q_4, and then extract the key feature points of each part for information fusion processing, as given in Eqs. (2) and (3):

$$Q_i^k = \left\{ \left(a_{ij}, b_{ij}\right)^k \right\} \begin{array}{l} i = 1,2,\cdots,t \\ j = 1,2,\cdots,m \end{array} \tag{2}$$

In Eq. (2), (a, b) represents the coordinates of key feature points; m represents the total number of features in each part; k represents the type of facial expressions; t represents the number of divided regions of the face; i represents the current i-th part area; and j represents the j -th feature key point in the i-th part.

$$G = \Phi\left(Q_1, Q_2, \cdots, Q_t\right) \tag{3}$$

In Eq. (3), Φ represents the feature fusion function and G is the result after feature fusion.

By using this method to train the deep learning model to achieve the effect of recognizing facial expressions, a schematic diagram of the existing deep learning framework for FER is shown in Fig. 3.

It can be seen from Fig. 3 that although the deep learning detection algorithm can accurately divide the various parts of the face $Q_1, Q_2, \cdots, Q_t$, and extracts the expression feature information $\left(a_{ij}, b_{ij}\right)$ between the various parts. But this method only works in each independent part, does not pay attention to the connotative connection between the parts of $Q_1, Q_2, \cdots, Q_t$, and ignores the internal connection between blocks. In other words, this method does not fully combine the implicit relationships between features to further realize FER.

Moreover, the key information of facial expression features extracted by this method is not all credible, which

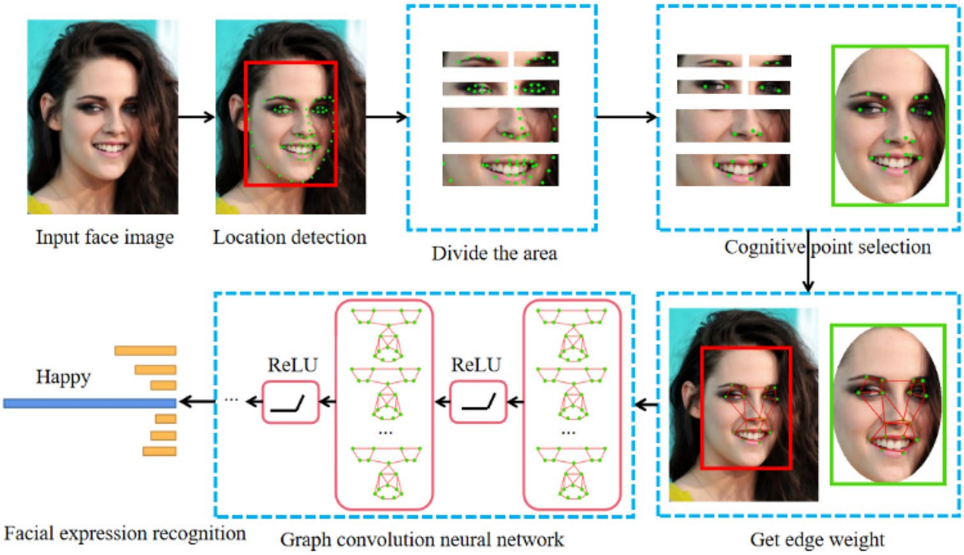
Input face image
Location detection
Divide the area
Cognitive point selection
Happy
ReLU
ReLU
...
Facial expression recognition
Graph convolution neural network
Get edge weight

Fig. 2 Schematic diagram of graph neural network model for facial expression recognition

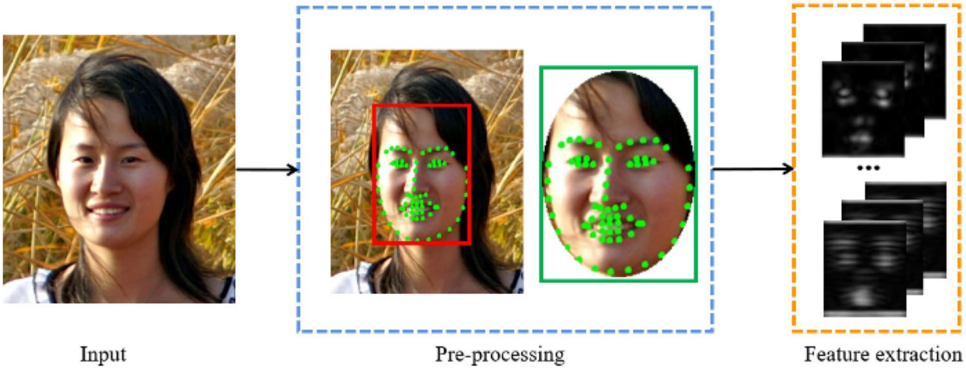

Fig. 3 Schematic diagram of facial expression recognition framework

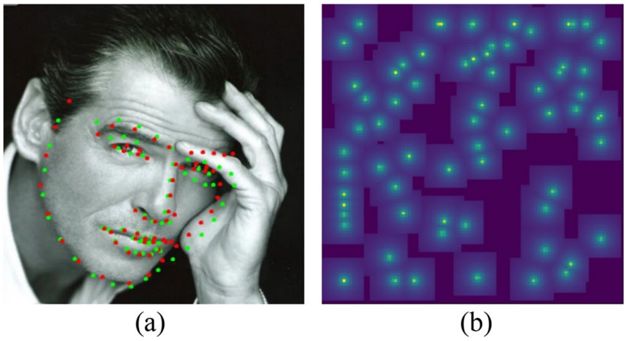

Fig. 4 Schematic diagram of key feature points of a human face

may cause the generalization ability of existing deep learning FER algorithms to weaken as shown in Fig. 4.

In Fig. 4a, the green dots represent the coordinates of the key feature points extracted by the detection algorithm, and the red dots represent the true coordinates of the key feature points. Figure 4b shows the heat map of the key feature points extracted by the detection algorithm. From Fig. 4a, it is clear that some green dots are farther from the red dots, and the deviation is large. From Fig. 4b, it can also be seen that some green dots have larger heat maps. All this shows that the key feature points extracted by traditional deep learning detection algorithms are not necessarily credible. If the model is constructed in this way (selecting all the key feature points), the accuracy of FER will be greatly affected. Moreover, when the model is used for FER, its features and recognition results are not well interpretable.

Based on some of the above drawbacks. We use the region division mechanism, and use the two key processes of local visual cognition and regional collaborative recognition to link the features of different regions, find the connotation relationship between originally independent parts, and enhance the generalization ability of the model. At the same time, we reduce some unreliable feature points through local uniform selection in the process of local visual cognition, and improve the accuracy of the model.

3.2 Facial expression recognition based on graph neural network

Visual cognition refers to the process by which people obtain knowledge or information through the visual system. Visual cognition is the foundation of human thought. When the human visual system is stimulated by the external environment, it first focuses its attention on the external information and performs local cognition by observing the salient features or key feature information points of the local information; then it performs regional collaborative association with the salient features of the observed information to identify the current information content. The process described above is the process of human visual cognition.

Following the logical process of visual cognition in the human brain, this paper divides the process of FER into two main stages: “local visual cognition” and “regional collaborative recognition”:

1) *Local visual recognition*: In the process of FER, it is necessary to recognize the current facial expression state k. First, the face Q to be judged needs to be divided into six parts: the upper-left area Q_1, the upper-right area Q_2, the upper-middle-left area Q_3, the upper-middle-right area Q_4, the middle area Q_5, and the lower area Q_6, as shown in Fig. 5 below:

   The division of the face area can be expressed by Eq. (4):

$$Q_i = \phi_i(Q)_{i=1,2,3,4,5,6} \tag{4}$$

   In Eq. (4), ϕ_i represents the function of extracting the face structure.

   Secondly, for each of the divided parts, the detection algorithm is used to cognitively locate the key feature points of each part, which can be expressed by Eq. (5):

$$Q_i^k = \left\{ \left(\alpha_{i1}, \beta_{i1}\right)^k, \left(\alpha_{i2}, \beta_{i2}\right)^k, \cdots, \left(\alpha_{in}, \beta_{in}\right)^k \right\} \tag{5}$$

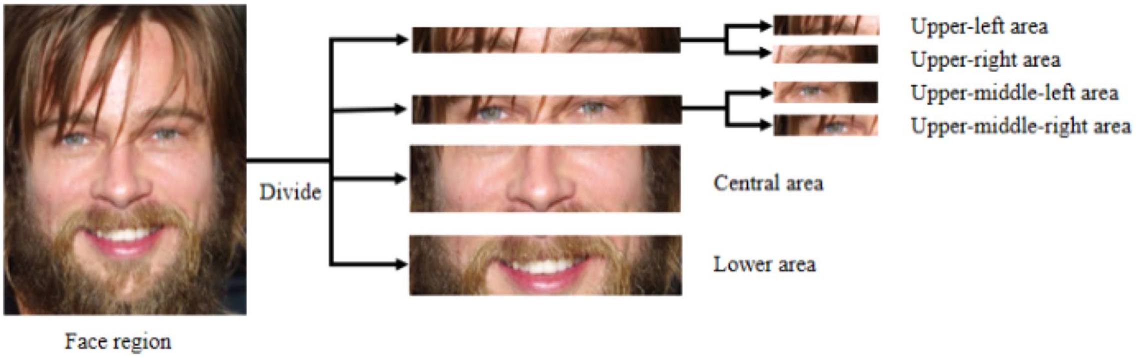

Fig. 5 Schematic diagram of face area division

In Eq. (5), i represents the i-th part; n represents the number of feature key points; and (α, β) refers to the key feature points of the i-th part.

Thirdly, according to the key points located in each part, we select n key points with strong cognitive credibility and make them evenly distributed in the face area (upper-left, upper-right, upper-middle-left, upper-middle-right, middle and lower). The credibility σ_j of the key points of current cognitive features can be obtained by Eq. (6):

$$\sigma_j = \begin{cases} \frac{\left|\rho - \left|\alpha_j - \alpha_j'\right|\right| \bullet \left|\rho - \left|\beta_j - \beta_j'\right|\right|}{\rho^2}, \left(\alpha_j - \alpha_j'\right) < \rho and \left(\beta_j - \beta_j'\right) < \rho \\ 0, otherwise \end{cases} \tag{6}$$

In Eq. (6), j refers to the j-th feature key point; $\left(\alpha_j', \beta_j'\right)$ is the real coordinate of the j-th feature key point; and ρ is the set coefficient. In this way, the credibility of each feature point in each part can be obtained, which can be expressed by Eq. (7):

$$P_i^k = \left\{\sigma_{i1}, \sigma_{i2}, \cdots, \sigma_{ij}, \cdots \sigma_{in}\right\} \tag{7}$$

Finally, according to the number of real key points in each partial area, the approximate ratio between the six areas is calculated to be $2 : 2 : 2 : 2 : 3 : 7$. According to the principle of equal proportions, the key points with strong credibility between each part are selected, and the corresponding credibility M can be obtained by the Eq. (8):

$$M_n^k = \Psi_i\left(P_i^k\right)_{i=1,2,3,4,5,6} \tag{8}$$

In Eq. (8), Ψ_i represents the function of selecting key points with strong credibility.

According to the mapping relationship between credibility and feature key points $f_i : M_i \to N_i$, the coordinate information N of the feature key points with strong credibility between each part is selected, as shown in Eq. (9):

$$N_n^k = \Upsilon_i\left(Q_i^k\right)_{i=1,2,3,4,5,6} \tag{9}$$

In Eq. (9), Υ_i represents the location extraction function with strong reliability for each part.

2) *Regional collaborative recognition*: In the process of regional collaborative recognition, it is necessary to analyse and synthesize the key points of the characteristics obtained from cognition and reorganize to make it systematized. For a given convolution kernel with a size of $Z \times Z$ and an input feature map f_{in} with the number of channels C, the position space s of a single channel can be expressed as Eq. (10) [42, 44]:

$$f_{ou}(s) = \sum_{h=1}^{Z} \sum_{w=1}^{Z} f_{in}(P(s, h, w)) \bullet w(h, w) \tag{10}$$

In the case of image convolution, the sampling function $P : G^2 \times G^2$ can also be expressed by Eq. (11):

$$P(s, h, w) = s + P'(h, w) \tag{11}$$

$w{:}G^2 \to R^C$ represents the weight function. To realize the "regional cooperative recognition", this paper redefines the weight function w. First, the distance d between the key points of the feature is calculated, which can be obtained by Eq. (12):

$$d_{(t-1,t)} = \sqrt{\left(x_t - x_{t-1}\right)^2 + \left(y_t - y_{t-1}\right)^2}, \left(x_t, y_t\right), \left(x_{t-1}, y_{t-1}\right) \in N_n^k \tag{12}$$

In Eq. (12), (x, y) represents the coordinates of the feature key point; and t represents the selected t-th feature key point.

Through centralized processing, calculate the distance between each feature key point and other feature key points and then add them to get $D_{(x_t,y_t)}$. It can be expressed as formula (13):

$$D_{(x_t,y_t)} = d_{(t-1,t)} + d_{(t-2,t)} + \cdots + d_{(0,t)} \tag{13}$$

The smaller the value of $D_{(x_i,y_i)}$, the closer the i-th cognitive key point is from the center. Therefore, starting from the key point with the smallest value of $D_{(x_i,y_i)}$, each cognitive feature key point is sorted in turn. Then determine a domain $B\left(N_i\right)$ for each selected cognitive feature key point N_i, where the mapping relationship can be expressed as $f_i : N_i \rightarrow B(N_i)$. Because there are Z labeled subsets in each field, we can get a mapping function $f_i : B(N_i) \rightarrow \{0, \cdots, Z-1\}$, and then map the neighborhood node to its corresponding subset. The weight function $w\left(N_i, N_j\right) : B(N_i) \rightarrow R^C$ of the key points of cognitive features N_i and N_j can be obtained by Eq. (14):

$$w\left(N_i, N_j\right) = w^{'}(f_i(N_j)) \tag{14}$$

The schematic diagram of the "regional collaborative recognition" process is shown in Fig. 6 as follows:

Based on the aforementioned human cognitive vision strategy, this paper constructs an undirected graph $G = (V, E)$ for each face image through the two processes of "local visual recognition" and "regional collaborative recognition". Each undirected graph has n cognitive nodes $(V = N = \{1, 2, \cdots, n\})$ and $E(E \subseteq V \times V)$ edges. Generally, the representation of an undirected graph depends on the adjacency matrix $W \in R^{|V|*|V|}$, and each element in W represents the weight of an edge, which is the connection strength between a pair of key feature nodes.

In order to address the shortcomings of the existing deep learning algorithms for FER, this work needs to adopt the key feature selection method based on the local visual recognition process and construct the edge of the regional collaborative recognition process to construct the undirected GNN, which is a more effective FER model. The model mainly implements two key processes (local visual recognition process and regional collaborative recognition process). First, the credibility between the key points of the facial features obtained by the detection algorithm and the corresponding key points of the actual features is calculated. Then, based on the credibility, the cognitive keypoint information of each face part (upper-left, upper-right, upper-middle-left, upper-middle-right, middle and lower) is selected, and the total distance of each cognitive keypoint to other cognitive keypoints for sorting is calculated; Finally, the weight function as the edge of the graph is constructed by the idea of 2D convolution operation to build the GNN model of visual cognitive strategy for FER. The specific process of FER algorithm based on the human GNN visual cognitive strategy is described in Algorithm 1.

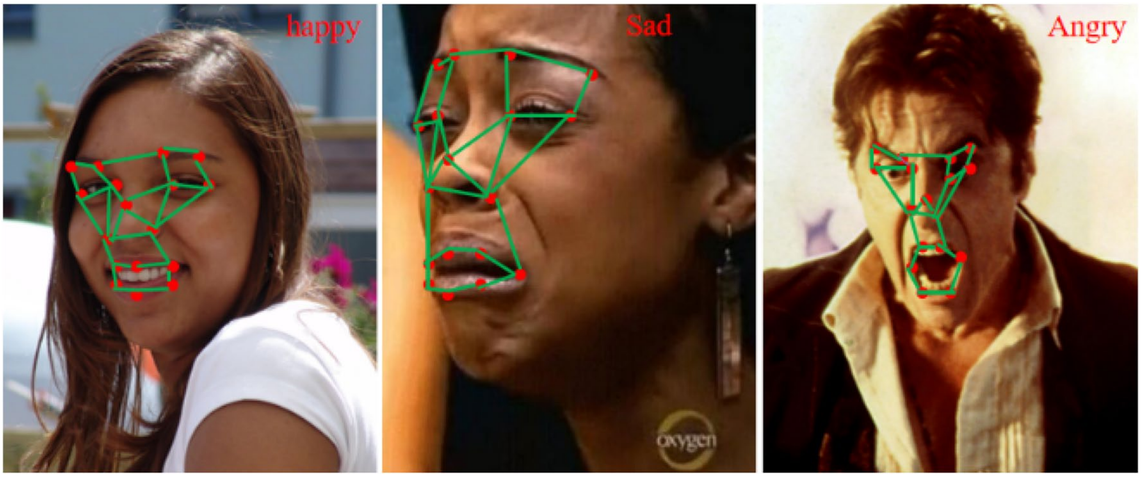

Fig. 6 Process diagram of "Regional Collaborative Recognition"

Algorithm 1: FER Algorithm based on Human Visual Cognitive Strategy of GNN

```
Input: Face image $img$;
Output: Facial expression state $k$;
Step1:
    Use the detection algorithm to locate the face area $Q$ and obtain the key points $\{(\alpha_{ij},\beta_{11})\}$ of face features in the face image $img$;
Step2:
    The face area $Q$ located in Step1 is divided into six parts: the upper-left area $Q_1$, the upper-right area $Q_2$, the upper-middle-left area $Q_3$, the upper-middle-right area $Q_4$, the middle area $Q_5$, and the lower area $Q_6$ based on the face structure features;
Step3:
  Repeat
     Repeat
          Calculate the credibility $\sigma_{i1},\sigma_{i2},\cdots,\sigma_{ij},\cdots\sigma_{in}$ of the key points of cognitive features through the key feature points of the face and the real key feature points located in Step1;
          Until Traverse all the key feature points of each part;
  Until All parts are traversed (upper-left, upper-right, upper-middle-left, upper-middle-right, middle and lower areas);
Step4:
    According to the credibility strength of the key points of cognitive features, select the key points of cognitive features $N_i$;
Step5:
    (1) Combine all the key points of cognitive features into $N$ sets;
    (2) Calculate the distance $d$ from each cognitive feature key point to other cognitive key points, and then sort each key point according to the centralization distance $D$;
    (3) By jointly cognizing the neighbor node set $B(N_i)$ of the node $N_i$, the face area $Q$ is divided into a fixed number of Z subsets. In this way, the mapping function $f_i{:}B(N_i)\to\{0,\cdots,Z-1\}$ can be obtained;
    (4) Map all neighbor nodes to other subsets, and the edge weight $w_i$ of each cognitive key can be obtained;
    (5) The GNN model is constructed by the key points $N_i$ selected by the local visual recognition process and the edge weights $w_i$ extracted by the regional cooperative recognition process.
Step6:
    Through the constructed GNN model, the facial expression state $k$ is obtained.
Return facial expression state $k$.
```

Time complexity analysis: The algorithm proposed in this paper is divided into two key processes: "partial visual recognition process" and "regional collaborative recognition process". In the process of local visual recognition, the facial expression area is first divided into $Q_1, Q_2, \dots, Q_i$ parts, each part has $n_1, n_2, \dots, n_i$ key feature points. The time complexity at this time is $O(1)$. Then, the credibility of the key feature points in each area of the face is calculated by Eq. (5). The time complexity at this time is $O(n_1), O(n_2), \dots, O(n_i)$. Finally, the ranking function is used to rank the credibility of the key feature points in each region. According to the result of ranking the credibility of key feature points,$N_1, N_2, \dots, N_i$ key cognitive features are selected in each area by using Eq. (8). The time complexity at this time is $O(N_1), O(N_2), \dots, O(N_i)$. Therefore, the time complexity of the local visual cognition process is $max(O(n_1), O(n_2), \dots, O(n_i))$. In the process of regional collaborative recognition, first the distance of each cognitive key point is calculated (the sum of the selected cognitive key points is $O(N)$, and the time complexity at this time is $O(\frac{N(N-1)}{2})$ (According to the nature of time complexity, $O(\frac{N(N-1)}{2})$ is equivalent to $O(N^2)$). Then, each key point is sorted according to the center distance D, and the time complexity at this time is $O(1)$. Finally, by mapping all neighbor nodes to other subsets, the edge weight of each cognitive key can be obtained, and the time complexity at this time is $O(N)$. Therefore, the time complexity of the regional collaborative recognition process is $O(N^2)$. According to the analysis of the time complexity of the local visual recognition process and the regional collaborative recognition process, the time complexity of the algorithm proposed in this paper is $O(N^2)$. Note: The time complexity of our algorithm does not consider the complexity of calling the function.

4 Experiment and analysis

4.1 Dataset and evaluation criteria

To verify the performance of the newly proposed model for FER, this paper uses the cross-validation method with the Helen dataset [45], the 300 W dataset, and the LFPW dataset [46].

The 300 W dataset and the LFPW dataset were released one after another at the 2013 International Computer Vision Conference. These two datasets are two types of data with marked key point information specifically used for automatic facial expression recognition. Most of the images were taken outdoors, while a small portion were taken indoors. The Helen dataset was proposed in 2012 and includes the changes in the background and facial expressions of the characters (e.g., changes in the poses of the subjects, changes in the background light, changes in the facial expressions, and so on). It is particularly important to note that before annotating the true location of the dataset, the data must be checked for anomalies in order to obtain high-quality annotated data results. Although the images in these three types of datasets all show human faces, there are no labels representing facial expressions. Therefore, this work mainly uses deep learning based FER algorithms and supplemented by a manual method to accurately assign these images with facial expressions to the seven types of tags/classes: angry, disgusted, fearful, happy, sad, surprised, and natural (inappropriate images are removed). The distribution of the number of images corresponding to the seven facial expressions in the above dataset is shown in Table 1.

Cross-validation is a method used to evaluate whether a trained model can be applied to another dataset with the same data structure. In this method, the original data is divided into two parts, one part is used as a training

Table 1 Distribution of instances in the training and validation sets of 300 W dataset, Helen dataset and LFPW dataset

	300 W dataset (600 sheets)	Helen dataset (2330 sheets)	LFPW dataset (1035 sheets)
Angry	108 (16.02%)	92 (3.90%)	45 (4.11%)
Disgust	11 (1.63%)	109 (4.62%)	105 (9.60%)
Fear	15 (2.23%)	53 (2.25%)	22 (2.01%)
Happy	316 (46.88%)	1031 (43.74%)	432 (39.49%)
Sad	42 (6.23%)	103 (4.37%)	36 (3.29%)
Surprise	19 (2.82%)	56 (2.38%)	16 (1.46%)
Neutral	163 (2.18%)	913 (38.74)	438 (40.04%)
Total	**674**	**2357**	**1094**

There may be multiple faces in some images, so the number of expressions may be greater than the actual number of pictures

dataset to train the model, and the other part is used as a test data set to evaluate the model. In this work, first, each data set (Helen dataset, 300 W dataset and LFPW dataset) is divided into 2 sets of data samples and these 2 sets of data samples are used for training respectively. Then another set of data samples is used as a test set for testing. Finally, the average recognition accuracy of these 2 models is used as the average performance standard of the model.

The experimental environment used in this paper is based on Pytorch1.4 under the Ubuntu16.04 system. The hardware environment consists of an Intel(R) Xeon(R) Gold5218 CPU, a GeForce RTX 1080Ti (1 GPU) with 30 GB memory, and a 1 T mechanical hard disk.

4.2 Ablation experiments on 300 W dataset

To investigate the behavior of the proposed method of local uniform selection, we performed an ablation study. From the theoretical point of view, the coordinate information of the key feature points of the face, the visual behavior of the local uniform selection of main features, and the random selection of key features are analysed, as shown in Fig. 7.

In the upper image in Fig. 7a, it is obvious that the reliability of some key feature points detected by the deep learning algorithm is low, that is to say, many feature points are unnecessary and may even have a bad impact on the final judgment. As shown in the upper image in Fig. 7b, when the key feature points with high reliability are selected by using the random selection method, the selected key features will not be evenly distributed, and the state of the entire facial expression cannot be effectively described. That is to say, the feature information of a relatively important part may be missed, such as eyebrows. Therefore, it is necessary to uniformly select the key feature points by dividing the face regions (upper left, upper right, middle left, middle right, upper middle, and lower middle), as shown in the upper row of images in Fig. 7c. Compared with Fig. 7a, the key feature points selected in Fig. 7c are more reliable and the model is more concise. Compared with Fig. 7b, c can well describe the state of facial expression, and there is no omission of a key part. It can also be seen from the heat map of the key feature points in the second row of the figure that (c) has a more concise and efficient representation. In

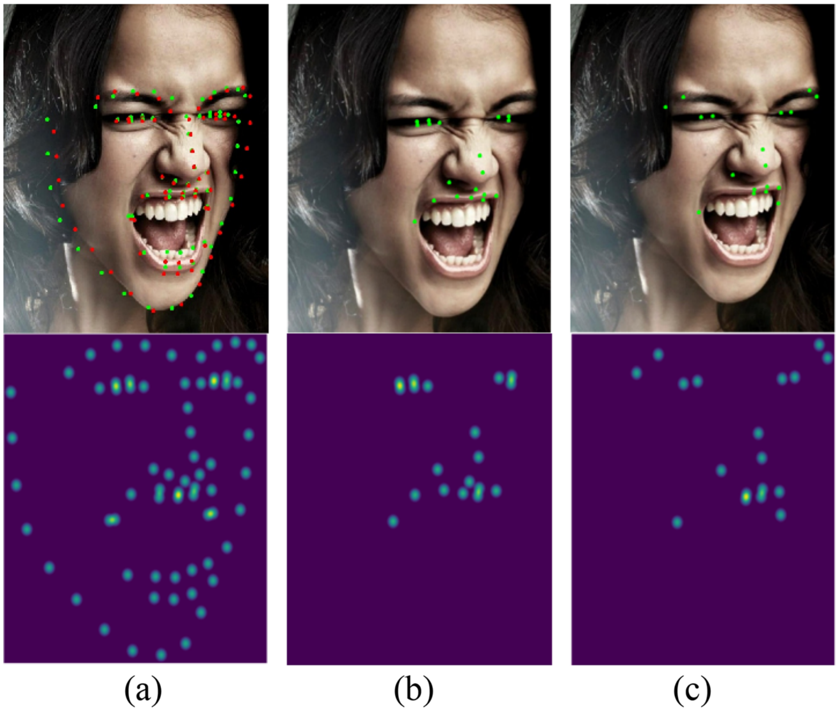

Fig. 7 Schematic diagram about random selection and uniform selection of key feature point information of the face

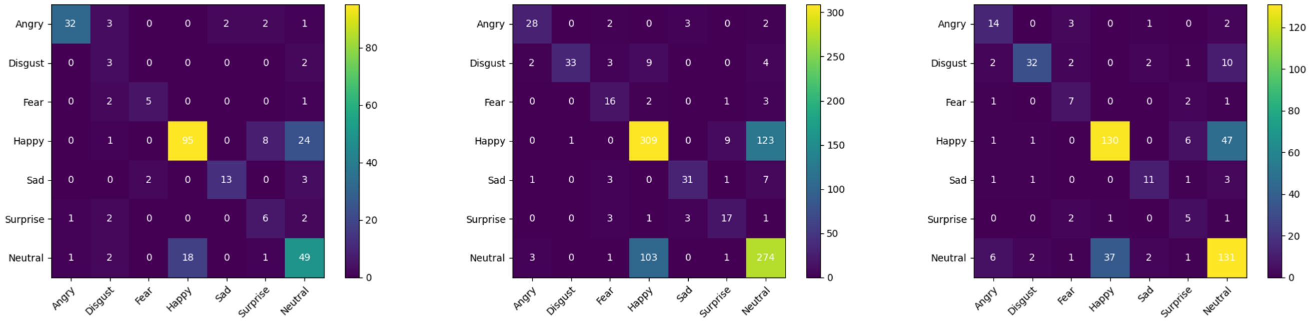

Fig. 8 The accuracy of the model constructed by random selection and uniform selection on the 300 W dataset

Table 2 The accuracy of the model constructed by random selection and uniform selection on the 300 W dataset

	Random model (%)	Uniform model (%)
Angry	47.22	67.59
Disgust	9.09	27.27
Fear	20.00	60.00
Happy	38.92	60.44
Sad	23.81	64.29
Surprise	21.05	47.37
Neutral	26.99	42.94
Total	**35.01**	**56.63**

other words, the uniform selection method can not only describe the entire face information, but also improve the characterization and generalization ability of GNN.

The two methods of locally uniformly selecting key feature point information and randomly selecting key feature point information were quantitatively analyzed using the 300 W dataset, as shown in Fig. 8. From Fig. 8, it can be seen that the GNN model created by the random key feature point selection method cannot effectively describe the facial expression features globally, resulting in poor model recognition. The GNN model constructed by uniformly selecting key feature points can recognize the relevant features of the face from the entire domain and helps to accurately recognize facial expressions.

It can be seen from Table 2 that the detection accuracy of the random key feature point selection method on seven different expressions is lower than that of the GNN model constructed by uniformly selecting key feature points, which indicates that uniformly selecting key feature points more effectively describes the overall facial features and helps to accurately identify facial expressions.

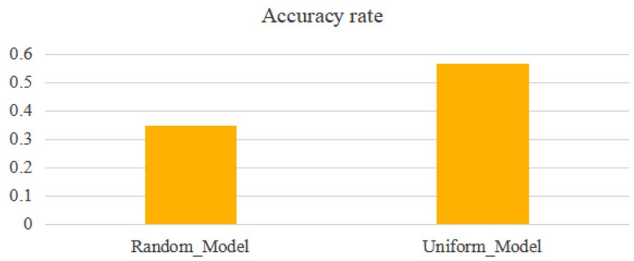

Fig. 9 Graph neural network model at 300 W, Helen, and LFPW verification concentrated heat map (The abscissa represents the classification result detected by the proposed algorithm, and the ordinate represents the correct classification result)

4.3 Quantitative analysis

To illustrate the feasibility of the model proposed in this article, the 300 W, Helen, and LFPW datasets are split into training and validation sets in a 7:3 ratio for verification. The heat map of the model for the 300 W, Helen, and LFPW validation sets is shown in Fig. 9.

The left side of Fig. 9 shows the heat map of the GNN model in the 300 W verification set, the middle of Fig. 9 shows the heat map of the GNN model in the Helen verification set, and the right side of Fig. 9 shows the heat map of the GNN model in the LFPW verification set. As can be seen from the three small images in Fig. 9, the main misclassification of facial expressions is the confusion between happy and natural expressions. In the 300 W, Helen, and LFPW verification sets, the percentages of happy expressions misclassified as natural expressions are 27.27%, 41.84%, and 32.87%, respectively, while the percentages of natural samples misclassified as happy expressions are 8.87%, 35.03%, 32.87%, and 25.87%. This high degree of confusion between happy and natural expressions is due not only to occlusion, postural changes, or other factors, but also to similar lip movements in the lower part of the face (mouth).

To further verify the effectiveness of the proposed GNN model based on visual cognitive strategies, this work compares the GNN model with traditional IACNN [47], Inception [32], VGG19 [48], Resnet18 [49], SAFC [50], and SCN [51] models on the 300 W dataset, Helen dataset, and LFPW dataset.

4.3.1 The 300 W dataset

The recognition rates of the proposed GNN-based FER model, the traditional IACNN model, the Inception model, the VGG19 model, the Resnet18 model, the SAFN model, and the SCN model are shown in Fig. 10. The recognition rate of the GNN model combined with the visual recognition strategy is 0.5663, while the recognition rates of the traditional deep learning models IACNN, Inception, VGG19, Resnet18, SAFC and SCN are 0.5036, 0.4814, 0.4763, 0.5136, 0.5546, and 0.5623, respectively. Compared with the traditional deep learning model, the model proposed in this paper increases the recognition rate of facial expressions by 12.45%, 17.64%, 18.90%, 10.26%, 2.11%, and 0.71%, respectively.

From Table 3, we can see the specific details of the comparison between our algorithm and the traditional algorithm. The recognition performance of facial expressions such as distinct, happy, surprise and neutral is the best, and the overall recognition result is higher than that of the traditional algorithm, which shows that our algorithm is indeed more effective than the traditional algorithm.

4.3.2 The Helen dataset

The recognition rates of the improved FER model based on GNN, the traditional IACNN model, the Inception model, the VGG19 model, the Resnet18 model, the SAFN model, and the SCN model are shown in Fig. 11. Based on the theory of visual cognition, this paper divides the FER process into two processes, “local visual recognition” and “regional collaborative recognition”, and finally reaches a FER rate of 0.5854. In contrast, the recognition rates of the traditional deep learning models IACNN, Inception, VGG19, Resnet18, and SCN are 0.5362, 0.5016, 0.5041, 0.5224, 0.5732, and 0.5769, respectively, indicating that the FER rate of the model proposed in this paper is higher than that of the traditional models by 9.18%, 16.71%, 16.13%, 12.06%, 3.71%, and 1.47%, respectively.

4.3.3 The LFPW dataset

The recognition rates of the new model proposed in this work, the traditional IACNN model, the inception model, the VGG19 model, the Resnet18 model, the SAFN model, and

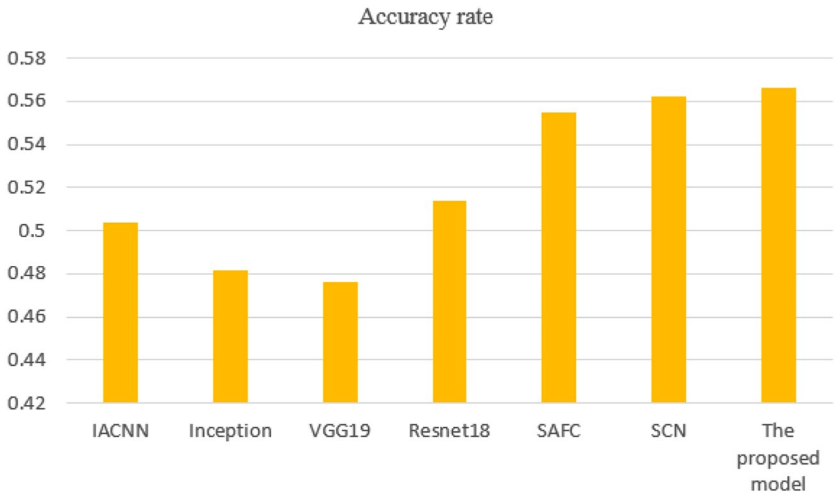

Fig. 10 The accuracy of the proposed model, IACNN, Inception, VGG19, Resnet18, SAFN and SCN models on the 300 W dataset

Table 3 The GNN model compared with the traditional IACNN [47], Inception [32], VGG19 [48], Resnet18 [49], SAFC [50], and SCN [51] models on the 300 W dataset(Among them, red, bold and italic respectively represent the first, second and third places)

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
IACNN [47]	**75.01%**	20.32%	37.50%	42.97%	61.11%	43.64%	*41.89%*	50.36%
Inception [32]	67.52%	19.32%	51.07%	40.06%	38.89%	42.73%	40.11%	48.14%
VGG19 [48]	72.46%	21.29%	25.64%	37.72%	55.56%	41.82%	39.30%	47.63%
Resnet18 [49]	70.44%	20.11%	***62.50%***	43.75%	52.20%	43.64%	40.71%	51.36%
SAFC [50]	***77.50%***	**24.17%**	55.66%	*57.42%*	*61.70%*	**43.81%**	41.39%	*55.46%*
SCN [51]	*74.16%*	*24.13%*	*58.36%*	**58.37%**	***65.30%***	*43.75%*	**42.33%**	**56.23%**
Ours	67.59%	***27.27%***	**60.00%**	***60.44%***	**64.29%**	***47.37%***	***42.94%***	***56.63%***

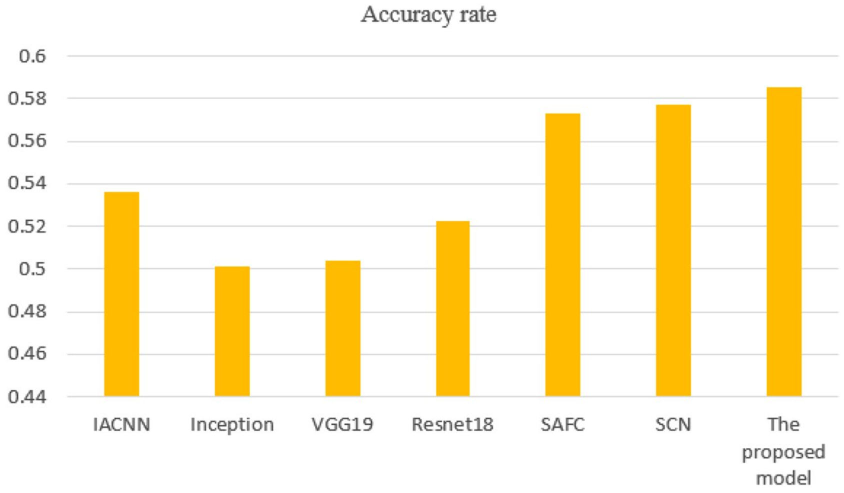

Fig. 11 The accuracy of the proposed model, IACNN, Inception, VGG19, Resnet18, SAFN and SCN models on the Helen dataset

the SCN model are shown in Fig. 12. By introducing visual perception theory into the face recognition algorithm, a FER rate of 0.5630 can be achieved, while the traditional FER models of IACNN, Inception, VGG19, Resnet18 and SCN can only achieve recognition rates of 0.5073, 0.4719, 0.4765, 0.4891, 0.5589 and 0.5596, respectively. Compared with the original face recognition model, the GNN model proposed in this work improves the recognition rates by 8.02%, 16.13%, 18.15%, 11.51%, 0.734%, and 0.61%, respectively.

4.4 Qualitative analysis

Automatic FER is a long-standing problem in computer vision. In this section, we verify the accuracy of the proposed GNN model in FER by testing several images with seven common facial expressions. The visualized result of FER based on the GNN model is shown in Fig. 13.

In Fig. 13, the upper row is a plot of visualization results from FER over a neural network constructed by uniformly selecting key feature points, and the lower row is a plot of visualization results from FER over a neural network constructed by randomly selecting key feature points. From the upper row of images in Fig. 13a, it can be seen intuitively that the lady's facial expressions are happy. However, it is not scientific to rely only on intuition. If you analyze and evaluate the model proposed in this article, the lady's facial expression can be considered happy with 95% probability. The model first learns and masters the information about the lady's facial expression in the image through the process of "local visual cognition", and then uniformly selects key cognitive feature points with high credibility as nodes of the graph, such as the raised corners of the mouth in the lower area, the slightly curved eyebrows in the upper area, and the curved eyes in the middle and upper area. Finally, in combination with the "regional collaborative recognition" process, the edge information between the key cognitive features is extracted as the edge of the graph to build a GNN model that achieves the effect of FER.

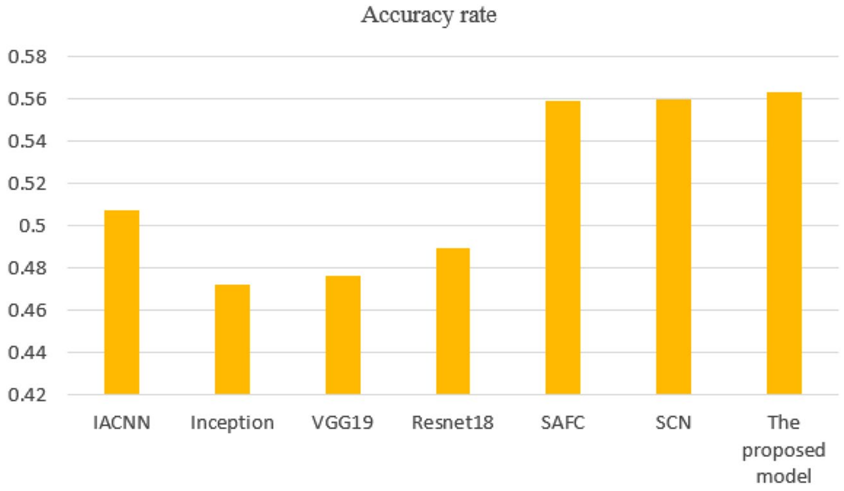

Fig. 12 The accuracy of the proposed model, IACNN, Inception, VGG19, Resnet18, SAFN, and SCN models on the LFPW dataset

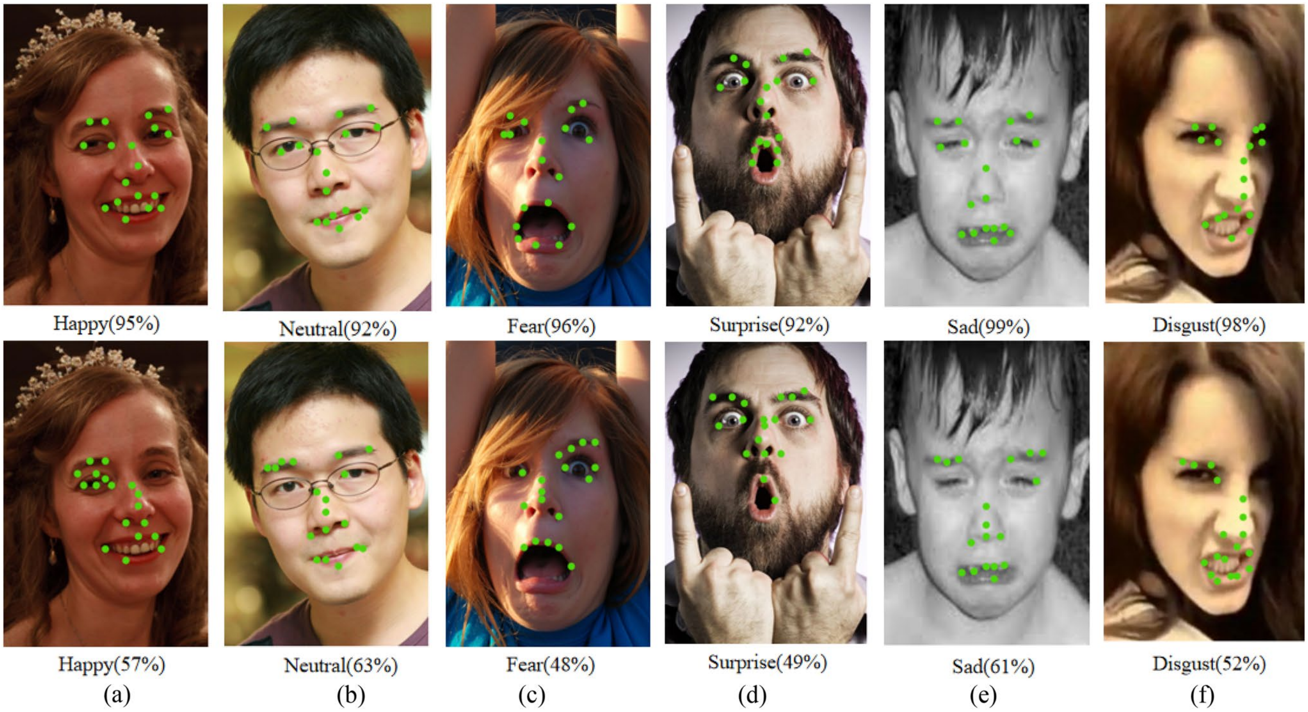

Fig. 13 The accuracy of the proposed model, IACNN, Inception, VGG19, Resnet18, SAFN and SCN models on the 300 W dataset

However, in the lower row of images in Fig. 13a, the random selection method is used, which cannot accurately describe the entire facial expression, so the accuracy of the model is only 57%. In the upper row of images in Fig. 13b, the facial features of the man do not vary significantly. Therefore, the proposed model classifies the man's facial expression as natural with a probability of 92%. However, in the lower row of Fig. 13b, the key feature points are randomly selected, so the constructed GNN model cannot describe the facial expression state well. In this way, the accuracy of the natural facial expressions is only 63%. In the upper row of Fig. 13c, the GNN model constructed in this paper mainly extracts obvious feature information such as the wide-open eyes in the middle and upper regions and the slightly downturned corner of the mouth in the lower region of the girl in the image. Through the joint analysis, it can be concluded that the facial expression of the girl can be considered as fear with 96% probability. The recognition rate of the randomly selected GNN model is only 48%. In the upper picture of Fig. 13d, the mouth in the lower area of a man's face corresponds to an “o” shape, the eyes are widened in the middle-upper area, and the eyebrows in the upper area are raised. The model proposed in this paper uses these salient features to conclude with a 92% probability that the man is pleasantly surprised. However, the upper row of images in Fig. 13d uses a random selection that cannot describe the mouth in the lower part of the face. This will make the recognition rate of the GNN model only 49%. In the upper image of Fig. 13e, the enhanced model learns the feature information about the boy's facial features through the process of “local visual recognition”, such as the downturned corners of the mouth in the lower area, the frowning brows in the upper area, and the drooping in the middle and upper areas. According to the “regional coordination” assessment, the boy's sadness is 99%. However, in the lower row of images in Fig. 13e, the random selection cannot describe the whole facial expression, so the recognition rate of the GNN model is only 61%. Similarly, for the proposed model, the girls in the upper image of Fig. 13f have a disgust level of 98%. However, the recognition rate using random selection in the lower row of Fig. 13f is only 52%.

It is not difficult to see from the Fig. 13 that the extraction of feature points of each region is more accurate after using the region division mechanism. Combining the two key processes of “local visual cognition” and “regional collaborative recognition”, a GNN model is constructed by using the idea of human visual cognition from local to global. This model can more accurately express facial expression with less feature point information, reducing the network parameters and improving the accuracy of expression recognition.

5 Discussion

This paper proposes a novel GNN model that accounts for the systematic process of FER in human visual cognition. The model uses local visual cognition to extract key

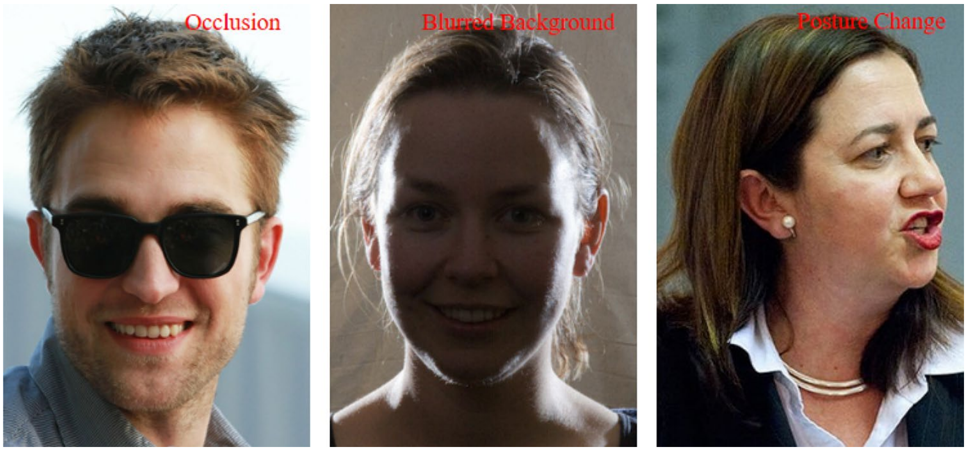

Fig. 14 Facial expression recognition challenges in occlusion, background blur, and posture changes

points of the face and performs regional collaborative fusion of the related information between the key points to realize FER. However, due to the complexity and variability of the dataset, the algorithm in this paper has some shortcomings, as shown in Fig. 14.

As can be seen in Fig. 14, the eyes in the left facial expression are completely occluded by the glasses, and the eyebrows are partially occluded state. The facial expression in the middle is blurred in the background. The posture of the face on the right has changed, and the other part of the facial expression information is not visible. We believe that the main reason for facial expression recognition errors is the failure of face detection when occlusion and facial posture change occur. At this time, the GNN model cannot correctly extract the information of key feature points of the face, so it cannot well construct the edge weights of key cognitive feature points. Therefore, the facial expression recognition results of the GNN model fail.

At the same time, due to the influence of artificial annotation and other reasons, happy and neutral expression are easily confused in the model. Therefore, in the future work, we must consider the following two aspects. On the one hand, we should consider the connotation relationship between facial expressions whose expression state changes every second or several seconds [52], and build a spatiotemporal neural network with stronger cooperative recognition ability; On the other hand, local expression information should be extracted to infer and predict the facial expression state of the domain. For example, only the mouth area is used to distinguish happy and neutral expressions, so that the difference between these two expressions can be amplified in the model.

6 Conclusion and future prospect

Existing FER methods focus on using deep learning to generate increasingly complex attention structures to extract facial features and achieve facial expression classification, and do not consider the internal connections between key facial feature points. Moreover, existing deep learning FER algorithms have serious cognitive difficulties, which further hinder the development of FER. Based on the above problems, this paper proposed a novel GNN model that can be applied to FER. The model firstly divides the human face into six different regions, extracts feature key points from the divided different regions evenly through “local visual cognition”, and then reveals the internal relationship between feature key points according to the concept of “regional cooperative recognition”, and constructs a GNN model to realize FER. By using cross-validation to compare with other related algorithms, this method not only guaranteed the effect of FER, but also expanded the application areas of neural network models. At the same time, it promoted the development of interpretability of cognitive science on GNN.

Our method also has the limitation that face detection may fail and some similar expressions are easily confused. In the follow-up work, we will try to eliminate these limitations by combining the time sequence of facial expression and the method of enhancing local feature information.

Acknowledgements The work was sponsored by Natural Science Foundation of Hunan Province with Nos. 2020JJ4434 and 2020JJ5368; Key Scientific Research Projects of Department of Education of Hunan Province with No. 19A312; Key Research Project on Degree and Graduate Education Reform of Hunan Province with No.2020JGZD025; National Social Science Foundation of China with No. AEA200013; Industry-Academic Cooperation Foundation of the Ministry of Education of China with No. HKEDU-CK-20200413-129.

Funding Open access funding provided by Western Norway University Of Applied Sciences.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Costa C, Antonucci F, Pallottino F et al (2011) Shape analysis of agricultural products: a review of recent research advances and potential application to computer vision. Food Bioprocess Technol 4(5):673–692. https://doi.org/10.1007/s11947-011-0556-0
2. Weinstein BG (2018) A computer vision for animal ecology. J Anim Ecol 87(3):533–545. https://doi.org/10.1111/1365-2656.12780
3. Fang W, Love P, Luo H et al (2020) Computer vision for behaviour-based safety in construction: a review and future directions. Adv Eng Inform 43:100980. https://doi.org/10.1016/j.aei.2019.100980
4. Kruger N et al (2013) Deep hierarchies in the primate visual cortex: what can we learn for computer vision? IEEE Trans Pattern Anal Mach Intell 35(8):1847–1871. https://doi.org/10.1109/TPAMI.2012.272
5. Yolcu G et al (2017) Deep learning-based facial expression recognition for monitoring neurological disorders. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 2017, pp 1652–1657, doi: https://doi.org/10.1109/BIBM.2017.8217907
6. Bouzakraoui MS, Sadiq A, Alaoui AY (2019) Appreciation of customer satisfaction through analysis facial expressions and emotions recognition. In: 2019 4th World Conference on Complex Systems (WCCS), Ouarzazate, Morocco, pp 1–5, doi: https://doi.org/10.1109/ICoCS.2019.8930761
7. Peng H, Yang R, Wang Z, Li J, He L, Philip SY, Zomaya A, Ranjan R (2021) Lime: low-cost and incremental learning for dynamic heterogeneous information networks. IEEE Trans Comput 71(3):628–642. https://doi.org/10.1109/TC.2021.3057082
8. Peng H, Zhang R, Li S, Cao Y, Pan S, Yu P (2022) Reinforced, incremental and cross-lingual event detection from social messages. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/tpami.2022.3144993
9. Xiao S, Wang L (2010) Reinterpretations of visual cognition from view of decomposition and synthesis. In: 2010 3rd International Congress on Image and Signal Processing, Yantai, China, pp 2828–2829, doi: https://doi.org/10.1109/CISP.2010.5647336
10. Summerfield C, Egner T (2009) Expectation (and attention) in visual cognition. Trends Cogn Sci 13(9):403–409. https://doi.org/10.1016/j.tics.2009.06.003
11. Green TM, Ribarsky W, Fisher B (2008) Visual analytics for complex concepts using a human cognition model. 2008 IEEE Symposium on Visual Analytics Science and Technology, Columbus, pp 91–98, doi: https://doi.org/10.1109/VAST.2008.4677361
12. Qiao H, Li Y, Li F, Xi X, Wu W (2016) Biologically inspired model for visual cognition achieving unsupervised episodic and semantic feature learning. IEEE Trans Cybern 46(10):2335–2347. https://doi.org/10.1109/TCYB.2015.2476706
13. Lathrop SD, Wintermute S, Laird JE (2011) Exploring the functional advantages of spatial and visual cognition from an architectural perspective. Top Cogn Sci 3(4):796–818. https://doi.org/10.1111/j.1756-8765.2010.01130.x
14. Shine JM, Breakspear M, Bell PT et al (2019) Human cognition involves the dynamic integration of neural activity and neuromodulatory systems. Nat Neurosci 22(2):289–296. https://doi.org/10.1038/s41593-018-0312-0
15. Qiao H, Xi X, Li Y, Wu W, Li F (2015) Biologically inspired visual model with preliminary cognition and active attention adjustment. IEEE Trans Cybern 45(11):2612–2624. https://doi.org/10.1109/TCYB.2014.2377196
16. Sakaguchi H, Utsumi A, Susami K, Kondo T, Kanbara M, Hagita N (2017) Analysis of relationship between target visual cognition difficulties and gaze movements in visual search task. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, pp 1423–1428, doi: https://doi.org/10.1109/SMC.2017.8122813
17. Micheli A (2009) Neural network for graphs: a contextual constructive approach. IEEE Trans Neural Networks 20(3):498–511. https://doi.org/10.1109/TNN.2008.2010350
18. Gnecco G, Gori M, Melacci S et al (2015) Foundations of support constraint machines. Neural Comput 27(2):388–480. https://doi.org/10.1162/NECO_a_00686
19. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS (2021) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32(1):4–24. https://doi.org/10.1109/TNNLS.2020.2978386
20. Gori M, Monfardini G, Scarselli F (2005) A new model for learning in graph domains. In: Proceedings 2005 IEEE International Joint Conference on Neural Networks, Montreal, pp 729–734, doi: https://doi.org/10.1109/IJCNN.2005.1555942
21. Krleža D, Fertalj K (2017) Graph matching using hierarchical fuzzy graph neural networks. IEEE Trans Fuzzy Syst 25(4):892–904. https://doi.org/10.1109/TFUZZ.2016.2586962
22. Shen Y, Li H, Yi S, et al (2018) Person re-identification with deep similarity-guided graph neural network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 486–504, doi: https://doi.org/10.1007/978-3-030-01267-0_30
23. Bongini M, Rigutini L, Trentin E (2018) Recursive neural networks for density estimation over generalized random graphs. IEEE Trans Neural Netw Learn Syst 29(11):5441–5458. https://doi.org/10.1109/TNNLS.2018.2803523
24. Rusek K, Suárez-Varela J, Almasan P, Barlet-Ros P, Cabellos-Aparicio A (2020) RouteNet: leveraging graph neural networks for network modeling and optimization in SDN. IEEE J Sel Areas Commun 38(10):2260–2270. https://doi.org/10.1109/JSAC.2020.3000405
25. Peng H, Zhang R, Dou Y, Yang R, Zhang J, Yu PS (2021) Reinforced neighborhood selection guided multi-relational graph neural networks. ACM Trans Inf Syst (TOIS) 40(4):1–46. https://doi.org/10.1145/3490181
26. Peng H, Li J, Song Y, Yang R, Ranjan R, Yu PS, He L (2021) Streaming social event detection and evolution discovery in heterogeneous information networks. ACM Trans Knowl Discov Data (TKDD) 15(5):1–33. https://doi.org/10.1145/3447585
27. Li S, Deng W (2020) Deep facial expression recognition: a survey. IEEE Trans Affect Comput 2:1. https://doi.org/10.1109/TAFFC.2020.2981446
28. Zhang L, Tjondronegoro D (2011) Facial expression recognition using facial movement features. IEEE Trans Affect Comput 2(4):219–229. https://doi.org/10.1109/T-AFFC.2011.13
29. Kyperountas M, Tefas A, Pitas I (2010) Salient feature and reliable classifier selection for facial expression classification. Pattern Recogn 43(3):972–986. https://doi.org/10.1016/j.patcog.2009.07.007

30. Cohen I, Sebe N, Garg A et al (2003) Facial expression recognition from video sequences: temporal and static modeling. Comput Vis Image Underst 91(1–2):160–187. https://doi.org/10.1016/S1077-3142(03)00081-X
31. Caifeng S, Shaogang G, McOwan PW (2005) Robust facial expression recognition using local binary patterns. IEEE International Conference on Image Processing, Genova, pp 2-370, doi: https://doi.org/10.1109/ICIP.2005.1530069
32. Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, pp 1–10, doi: https://doi.org/10.1109/WACV.2016.7477450
33. Zeng N, Zhang H, Song B, Liu W, Li Y, Dobaie AM (2018) Facial expression recognition via learning deep sparse autoencoders. Neurocomputing 273(17):643–649. https://doi.org/10.1016/j.neucom.2017.08.043
34. Yurtkan K, Soyel H, Demirel H (2015) Entropy driven feature selection for facial expression recognition based on 3-D facial feature distances. In: 2015 23nd Signal Processing and Communications Applications Conference (SIU), Malatya, pp 2322–2325, doi: https://doi.org/10.1109/SIU.2015.7130344
35. Khan S, Chen L, Yan H (2020) Co-clustering to reveal salient facial features for expression recognition. IEEE Trans Affect Comput 11(2):348–360. https://doi.org/10.1109/TAFFC.2017.2780838
36. H Jun, C Jian-Feng, F Ling-Zhi, H Zhong-Wen (2015) A method of facial expression recognition based on LBP fusion of key expressions areas. The 27th Chinese Control and Decision Conference (2015 CCDC), Qingdao, pp 4200–4204, doi: https://doi.org/10.1109/CCDC.2015.7162668
37. Shojaeilangari S, Yau W, Nandakumar K, Li J, Teoh EK (2015) Robust representation and recognition of facial emotions using extreme sparse learning. IEEE Trans Image Process 24(7):2140–2152. https://doi.org/10.1109/TIP.2015.2416634
38. Li Y, Wang S, Zhao Y, Ji Q (2013) Simultaneous facial feature tracking and facial expression recognition. IEEE Trans Image Process 22(7):2559–2573. https://doi.org/10.1109/TIP.2013.2253477
39. Mohan K, Seal A, Krejcar O, Yazidi A (2021) Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. IEEE Trans Instr Meas 70:1–12. https://doi.org/10.1109/TIM.2020.3031835
40. Lee J, Kim S, Kim S, Sohn K (2020) Multi-modal recurrent attention networks for facial expression recognition. IEEE Trans Image Process 29:6977–6991. https://doi.org/10.1109/TIP.2020.2996086
41. Meng Z, Liu P, Cai J, Han S, Tong Y (2017) Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, pp 558–565, doi: https://doi.org/10.1109/FG.2017.140
42. André TL, de Edilson A, Alberto FDS, Thiago O-S (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn 61:610–628. https://doi.org/10.1016/j.patcog.2016.07.026
43. Sun N, Li Q, Huan R et al (2019) Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recogn Lett 119:49–61. https://doi.org/10.1016/j.patrec.2017.10.022
44. Peng H, Li J, Gong Q, Song Y, Ning Y, Lai K, Yu PS (2019) Fine-grained event categorization with heterogeneous graph convolutional networks. Proc IJCAI 8:3238–3245. https://doi.org/10.24963/ijcai.2019/449
45. Belhumeur PN, Jacobs DW, Kriegman DJ, Kumar N (2013) Localizing parts of faces using a consensus of exemplars. IEEE Trans Pattern Anal Mach Intell 35(12):2930–2940. https://doi.org/10.1109/TPAMI.2013.23
46. Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M (2013) 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: 2013 IEEE International Conference on Computer Vision Workshops, Sydney, pp 397–403, doi: https://doi.org/10.1109/ICCVW.2013.59
47. Zhang C, Wang P, Chen K, Kämäräinen J-K (2017) Identity-aware convolutional neural networks for facial expression recognition. J Syst Eng Electron 28(4):784–792. https://doi.org/10.21629/JSEE.2017.04.18
48. Simonyan K, Andrew Z (2022) Very deep convolutional networks for large-scale image recognition, ICLR 2015 . International Conference on Learning Representations
49. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp 770–778, doi: https://doi.org/10.1109/CVPR.2016.90
50. Li S, Deng W (2020) A deeper look at facial expression dataset bias. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2020.2973158
51. Wang K, Peng X, Yang J, Lu S, Qiao Y (2020) Suppressing uncertainties for large-scale facial expression recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, pp 6896–6905, doi: https://doi.org/10.1109/CVPR42600.2020.00693
52. Peng H, Li J, Gong Q, Ning Y, Wang S, He L (2020) Motif-matching based subgraph-level attentional convolutional network for graph classification. Proc AAAI Conf Artif Intell 34(04):5387–5394. https://doi.org/10.1609/aaai.v34i04.5987

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.