Neurocomputing 462 (2021) 320–330

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

SG-DSN: A Semantic Graph-based Dual-Stream Network for facial expression recognition

Yang Liu [a,b], Xingming Zhang [a,*], Jinzhao Zhou [a], Lunkai Fu [a]

[a] School of Computer Science and Engineering, South China University of Technology, China
[b] Center for Machine Vision and Signal Analysis, University of Oulu, Finland

ARTICLE INFO

Article history:
Received 16 December 2020
Revised 6 May 2021
Accepted 4 July 2021
Available online 7 July 2021
Communicated by Zidong Wang

Keywords:
Facial expression recognition
Affective computing
Graph representation
Graph convolutional attention block
Semantic relationship

ABSTRACT

Facial expression recognition (FER) is a crucial task for human emotion analysis and has attracted wide interest in the field of computer vision and affective computing. General convolutional-based FER methods rely on the powerful pattern abstraction of deep models, but they lack the ability to use semantic information behind significant facial areas in physiological anatomy and cognitive neurology. In this work, we propose a novel approach for expression feature learning called Semantic Graph-based Dual-Stream Network (SG-DSN), which designs a graph representation to model key appearance and geometric facial changes as well as their semantic relationships. A dual-stream network (DSN) with stacked graph convolutional attention blocks (GCABs) is introduced to automatically learn discriminative features from the organized graph representation and finally predict expressions. Experiments on three lab-controlled datasets and two in-the-wild datasets demonstrate that the proposed SG-DSN achieves competitive performance compared with several latest methods.

1. Introduction

Facial expression recognition (FER) has become an attractive research area in recent years, as it plays a significant role in many applications such as face animation [1] and medical diagnosis [2]. One key challenge of implementing effective FER is to capture discriminative expression information from static images or video sequences. Previous studies mainly depend on hand-craft feature design or automatic feature learning followed by classifier construction [3,4]. However, these methods generally handle local and holistic expression cues in the view of classic image processing, without considering latent semantic information. In this work, we aim to develop a principled and effective method that combines the facial appearance and geometric information with their semantic relationships and leverage it for FER.

In the last few years, the convolutional neural network (CNN) has achieved great improvement in the FER rate and eliminated the tedious design of the hand-craft feature [5,6]. Recently, researchers attempt to optimize the feature learning process that can utilize spatial or temporal expression information to enhance FER performance [7,8]. Nevertheless, most of these methods simply regard facial expressions as dynamic variations of several key parts. The capability is limited since they do not explicitly consider spatial or co-occurring relationships among these facial areas, which are crucial for understanding facial expression according to physiological anatomy and cognitive neurology studies [9–11].

To move beyond the above drawbacks, we need a novel approach that can automatically learn patterns contained in key facial parts as well as their semantic relationships. For local texture features, it is the strength of classic image processing technology to capture discriminative expression information from different dimensions. For semantic relationships, one feasible way is to exploit the graph structure based on facial landmarks to represent faces, which is more consistent with facial muscle anatomical definition [12,13]. Yet, the non-grid structure of graphs makes it difficult to use standard deep models like CNNs. Currently, graph neural networks (GNNs) have received increasing attention and have successfully been generalized to lots of computer vision tasks, such as image-text matching [14] and human action recognition [15]. Thus, how to exploit a graph to encode both spatial and semantic information as well as how to implement GNNs to learn discriminative features from the graph representation are two major problems in graph-based FER.

In this paper, we design a graph representation of facial expression followed by an extended GNN, called Semantic Graph-based Dual-Stream Network (SG-DSN). The proposed graph representation is generated based on facial landmarks, where each node indi-

* Corresponding author.
E-mail address: cszxm@scut.edu.cn (X. Zhang).

https://doi.org/10.1016/j.neucom.2021.07.017
0925-2312/

cates a local patch around one landmark. The semantic connections among every node pair are initialized as edges of the facial graph. Then, the dual-stream network (DSN) integrates both appearance and geometric variations as well as their semantic relationships embedded in the organized facial graph to learn effective features and classify facial expressions.

The main contributions of this work summarize in four aspects:

- A graph representation for modeling facial expressions is generated, which consists of reasonable landmarks and semantic connections based on prior knowledge in physiological anatomy and cognitive neurology.
- A variety of local feature extraction methods and indexing initialization strategies are designed and evaluated for effective description of node attributes and edge attributes respectively.
- A dual-stream network is built by stacking graph convolution layers with attention blocks to learn discriminative features, which can integrate both local variations and their semantic relationships for expression prediction.
- On five public datasets, the proposed SG-DSN achieves remarkable performance against previous FER methods.

2. Related work

2.1. Semantic expression representation

An effective facial representation can not only focus on the critical facial areas but also eliminate the useless information caused by background noises or facial organ deformations. Generalizing facial semantic information to describe expressions is an emerging topic in FER research. Previous studies usually cropped images based on basic facial components (e.g. eyes, nose, and mouth) and then captured the local texture and spatial relationships from facial patches. Zhang et al. [16] decomposed facial landmarks into different parts to extract dynamic semantic geometric information from facial morphological variations, which complemented the static appearance features. Ye et al. [17] proposed a region-based deep model to fuse semantic information among different levels of receptive fields within valuable and unified patches. Recently, several graph-based methods have been designed for more effective facial representation, which systemically modeled semantic information in a static sense or a function of time. Liu et al. [18] built an action units graph that encoded both appearance and geometric expression information and co-occurring action unit relations to achieve effective representation for FER. Li et al. [19] presented a semantic relationship embedded representation learning framework through structured knowledge-graph to generate enhanced facial representation. Zhong et al. [20] utilized a graph structure to represent facial expressions for removing useless information and depicting geometric changes within different facial expressions. Zhang et al. [21] introduced a context-aware affective graph to extract context elements for discrete emotion inferring and achieved higher performance than previous methods. However, most of the studies regarded the graph as an independent geometric branch outside the facial appearance that results in limited performance. And some methods also demand temporal information when building facial graphs, which cannot be implemented on the latest large-scale FER dataset. In this paper, a graph representation is constructed to jointly model key facial variations and their semantic relationships based on static images.

2.2. Graph neural network

GNNs are widely used in many artificial intelligence tasks, due to the ability of feature learning for graph structure data. In the field of computer vision, GNNs can be sorted into two categories with different processing thoughts: the input-based and the network-based. The first type attempts to transform graph structure data into forms that can be trained by standard deep models. Such et al. [22] presented a Graph-CNN learning framework for image classification that could process graph data while maintaining the advantage of standard CNNs. Walecki et al. [23] provided an auto-encoder to fuse local expression confidence values and then predicted facial expressions. In [24], a structured deep network was put forward to model graph inputs and generate complex feature representations simultaneously for expression intensity estimation. By contrast, network-based approaches aim to build specific neural networks that are more suitable for graph structure data than input-based ones. Zhou et al. [25] proposed a spatial–temporal graph convolutional network to learn both spatial and temporal patterns from graph data and had proved its applicability to FER. Li et al. [19] introduced a gated graph neural network (GGNN) in a multi-scale CNN framework for propagating node information to improve expression representation. Zhang et al. [21] proposed a graph-based reasoning network to learn the affective relationship during the back-propagation process and outperformed the baseline methods. Zhong et al. [20] exploited a bidirectional recurrent neural network (BRNN) to iterate each node on a facial graph for the extraction of appearance and geometric patterns. Different from the above methods that independently considering the relational dependencies among facial areas, we propose a dual-stream GNN to learn spatial features and semantic relationships simultaneously. In addition, an attention module like [26] is employed to explicitly enhance the semantic relational reasoning of facial variations, which improves the effectiveness and interpretation of previous methods.

3. Proposed method

Recent physiological and psychological studies have revealed that different expressions can be recognized by perceptual factors in key facial areas and specific semantic information in facial context [9,10]. Inspired by this prior knowledge, a novel graph-based FER method is proposed in this section. Specifically, we firstly demonstrate the processes of the local feature extraction and the semantic relationship initialization. The two outputs are defined as the node and edge attributes separately to construct our graph representation. Next, we design a dual-stream network (DSN) by stacking graph convolutional attention blocks (GCABs) to learn features from the facial graph for effective FER.

The pipeline of our SG-DSN is illustrated in Fig. 1. An input image is first preprocessed with face detection and rotation correction. Next, local texture features are extracted based on detected and computed landmarks, while edge indexes are initialized at the same step. After the facial graph is generated, it is then transformed as the input of the DSN for final facial expression prediction.

3.1. Facial graph representation

Based on the theory of cognitive neuroscience about face perception, human beings use a dual-system model to process and recognize facial information: analytic processing and holistic processing [27]. Specifically, analytic processing obtains corresponding multi-dimensional cluster features by analyzing local areas of the face, while holistic processing aims to generate a holistic representation to perceive the overall structure among critical facial parts [27,28]. Therefore, in order to formally exploit above intrinsic properties of facial expressions, a reasonable way is to introduce the graph structure for expression modeling [18,20,21]. When one face is described by a graph, not only the scattered facial

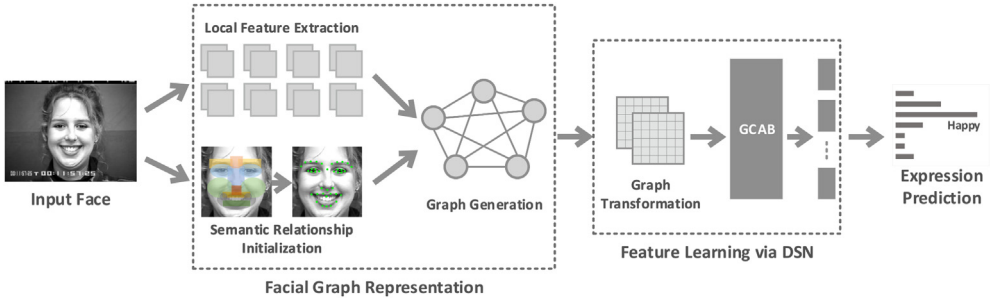


Fig. 1. The pipeline of our SG-DSN method.

changes can be integrated, but also the expression semantic relationship can be embedded to form a local-to–holistic facial representation. Therefore, a graph representation is designed by the local feature extraction and the semantic relationship initialization in this section.

3.1.1. Local feature extraction

Local feature extraction aims to focus on texture changes in specific facial areas among different expressions. Generally, facial landmarks follow facial muscle anatomy that can be further used to locate target facial patches. In this work, 68 facial landmarks are firstly detected and then 30 (including 17 of the external outline, five of the nose contour, and eight of the mouth) of them are discarded due to their non-saliency of expression. This landmark selection is based on facial topology and FACS definition, which is widely used in previous studies [13,20,25]. In addition, to cover the texture of the forehead area where key action units (AU) may activate (e.g., AU6: brow lowerer; AU9: nose wrinkler), two additional landmarks are calculated based on existing ones by:

$$l_{p'} = 0.5 \times l_{22} + 0.5 \times l_{23}, \tag{1}$$

and

$$l_{p''} = 2 \times l_{p'} - l_{28}, \tag{2}$$

where l_i denotes the landmark coordinate. As shown in Fig. 2, we get a total of 40 reasonable facial landmarks. The contribution of the landmark selection and addition is evaluated in Appendix A.

Next, the local texture information around each facial landmark $p \in P$ is extracted. We assume that the result of analytic processing can be achieved by using different feature detectors. Considering the patch size of landmark neighbors is usually small, we propose two methods to conduct local feature extraction.

One way is to use fused classic features. Here, Gabor filters and HOG descriptors are applied for effectiveness and convenient computation. Fig. 3 presents the overall framework of the process. Specifically, we set

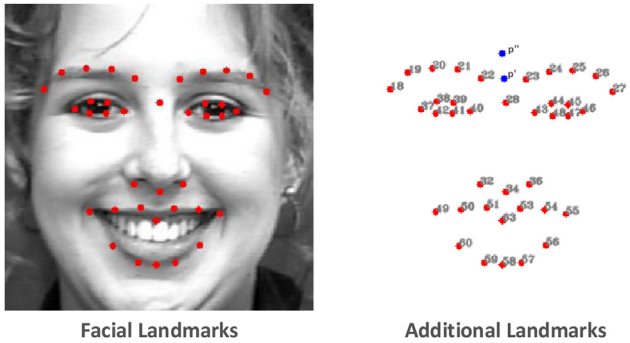


Fig. 2. Detection and calculation of facial landmarks.

$\theta = 0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4$,
$\lambda = 4, 4\sqrt{2}, 8, 8\sqrt{2}, 16$ to generate 40 sets of Gabor vectors, where θ and λ are two important parameters that indicate wavelength and direction of the filter respectively. These vectors are then fused averagely in eight directions to output the local Gabor feature f_p^g attached to landmark p. For HOG descriptors, nine feature maps are generated based on the gradient amplitude and orientation at each pixel. Next, the obtained feature maps are applied to encode local feature vector f_p^h about the area of landmark p through corresponding feature channels. Effects of different Gabor kernel sizes and different cell and block numbers of HOG are evaluated in Section 4.2.1.

Both Gabor and HOG feature vectors are then concatenated to generate the complete local texture feature. Thus, the local texture feature to the neighborhood of landmark p can be formulated by:

$$f_p = Concat(f_p^g, f_p^h), \tag{3}$$

where $Concat(\cdot)$ is the concatenation function. Since the two types of detectors have the same order of magnitude (640 and 324 for Gabor and HOG respectively in this work), powerful performance can be achieved by directly feature concatenation. But when choosing other kinds of local texture detectors, feature balancing may be an important option.

Furthermore, we also attempt to design a lite-CNN to learn local texture features following the intuition behind well-known models like VGG-Face [29], because the CNN-based descriptor has proven its effectiveness in previous work [6], even if it is more time-consuming. As shown in Fig. 4, our lite-CNN has a five-layer network architecture with an image patch size of 16×16 as input, where two convolutional layers and two max-pooling layers are alternatively stacked, followed by a fully connected layer. More concretely, the convolutional operation uses 3×3 size kernels and is implemented in a stride of 1 without padding, while the kernel size of max-pooling layers is 2×2. And the output feature maps are fully connected to generate the final local texture feature vector f_p. Note that the input size is the same as that used in the above fused feature to balance the need for performance and fair comparison. And before feeding all the local patches, the lite-CNN will first be pre-trained like other CNN-based FER methods [19]. The performance evaluation between these two local feature extraction methods is conducted in Section 4.2.2.

3.1.2. Semantic relationship initialization

The next key part is to simulate the procedure of holistic processing. Since every facial expression consists of a combination of local facial variations, how to model this kind of relationship is the other critical job. Fig. 5 shows an example of how to build the semantic relationship of *happiness*. The blue lines in the left sub-figure indicate possible semantic dependencies. Specifically, edge connections are first initialized based on prior knowledge,

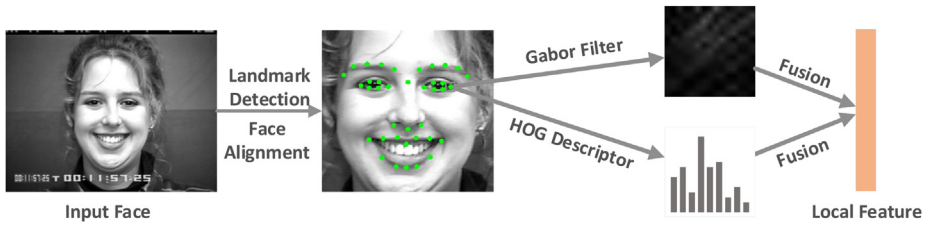


Fig. 3. The framework of local feature extraction via Gabor and HOG.

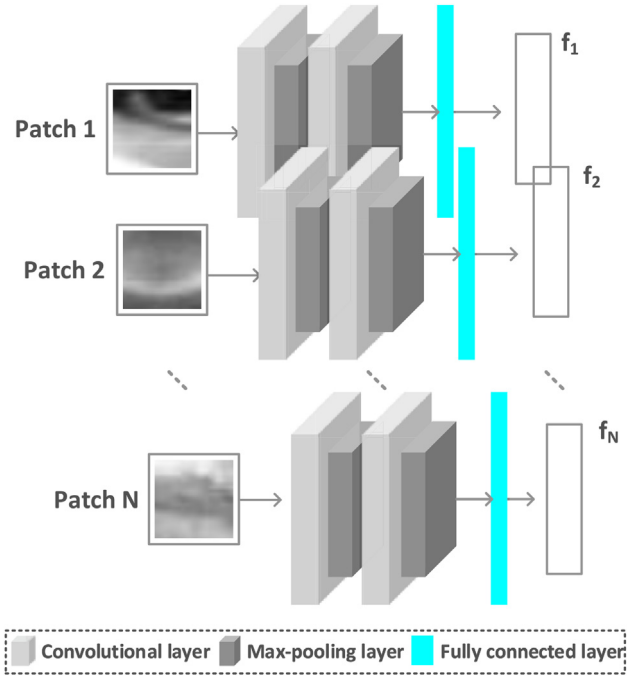


Fig. 4. The framework of local feature extraction via lite-CNN.

including FACS and facial topology [12,13,25], to establish semantic dependencies among facial changes. Furthermore, edge indexes are calculated to introduce attributes of these pre-defined links. In this subsection, we design two edge indexing strategies: Euclidean distance and Hop distance.

Euclidean distance: For any two nodes (landmarks) p, q in the node set P, their edge indexes can be calculated by:

$$s_{p,q} = \frac{\|l_p - l_q\|_2}{D_{eye}}, \tag{4}$$

where l_p, l_q are the coordinates of landmarks p, q, and D_{eye} is the inner-eyes distance which is used for normalization of scale diversity.

Hop distance: Let A indicates the adjacency matrix of the initialized graph, it is easy to compute the nearest hop matrix A', in which A'_{pq} denotes the shortest hop distance between any two nodes p, q. Considering a semantic facial action may occur in a joint region composed of several adjacent nodes, a node may also interact with the neighbors of its connected node. So the hop distance can be formulated as their edge indexes by:

$$s_{p,q} = \begin{cases} A'_{pq}, & \text{if } A'_{pq} \leqslant B, B \in \{1,2\} \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where B is defined as the receptive threshold for all the connected node pairs, that is B denotes the max acceptable hop distance on the graph. In practice, we set $B = 2$ for a trade-off between efficiency and relevance.

However, either Euclidean distance or Hop distance is only initial edge attributes and needs dependency reasoning to learn semantic features. Besides, the location of landmarks is equally important for graph node representations. Thus, landmark coordinates are taken together to form the global geometric cues of the facial graph. Similarly, the effects of these two strategies and their settings are compared in Section 4.2.3.

3.1.3. Graph generation

After the two steps above, we can present the definition of our graph representation. Definition 1 explains the details of the facial graph in this work.

Definition 1. Let $G = (V, E)$ denotes a facial graph: $\forall v_i \in V$ is a region near one facial landmark; $\forall e_j \in E$ is a 2-element subset of V that represents any edge existing in graph. $P^v = \{l_i | 1 \leqslant i \leqslant |V|\}$ denotes the landmark coordinates. $F^v = \{f_i | 1 \leqslant i \leqslant |V|\}$ is the node attribute set, and f_i represents the extracted local texture feature of the corresponding facial patch v_i, and $|V|$ is the number of

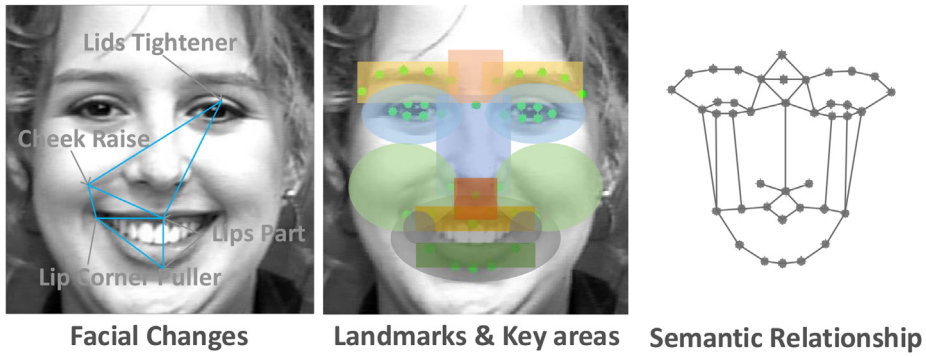


Fig. 5. An example of semantic relationships in *happiness*.

landmarks in V. $F^e = \{s_j | 1 \leqslant j \leqslant |E|\}$ is the edge attribute set, s_j indicates the semantic distance of edge e_j, and $|E|$ is the number of edges in E.

Note that the obtained local texture features and initial edge indexes are taken as the node attribute set F^v and the edge attribute set F^e for graph generation respectively.

There are two advantages of the proposed facial graph:

1. The facial graph describes expressions from a formalized local-to-holistic view and keeps the latent semantic information by using graph structure;
2. The facial graph integrates both appearance and geometric information of expressions that can provide sufficient cues.

In the next subsection, we introduce a graph-based neural network to learn effective features from the generated facial graph for enhanced FER.

3.2. Dual-stream graph network for FER

Different from classic CNNs that have the input with grid structures, GNNs can manage graph structure data and maintain the effectiveness of the convolution procedure. In this subsection, a GNN based on graph convolution and attention module is designed to process the above generated facial graph for expression feature learning.

3.2.1. Graph convolutional attention block

Given a graph G with nodes and their representations, the idea of graph convolution is designed in the Fourier domain by the multiplication of a signal $x \in \mathbb{R}^N$ with a filter $g_\theta = \mathrm{diag}(\theta)$ parameterized by $\theta \in \mathbb{R}^N$ as follows:

$$g_\theta \cdot x = U g_\theta U^\top x, \tag{6}$$

where U is the matrix of eigenvectors of the normalized graph Laplacian $L = I_N - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = U \Lambda U^\top$ (D indicates the degree matrix and A is the adjacency matrix), with a diagonal matrix of its eigenvalues Λ. Specifically, the graph convolutional layer (GCL) generalizes the definition to a signal $X \in \mathbb{R}^{N \times C}$ with C input channels (C is the vector dimension of each vertex attribute in our work) and K filters for feature maps as follows:

$$Z = \bar{D}^{-\frac{1}{2}} \bar{A} \bar{D}^{-\frac{1}{2}} X \Theta, \tag{7}$$

where $\bar{A} = A + I_N$ and $\bar{D}_{ii} = \Sigma_j \bar{A}_{ij}$, Θ is a matrix of filter parameters and $Z \in \mathbb{R}^{N \times K}$ is the convolved signal matrix. The GCL can be trained on a specific structure by learning filters based on the eigendecomposition of the graph Laplacian that is suitable for our proposed facial graph.

In addition, to further extract the semantic relationship among the facial graph, we design the graph convolutional attention block (GCAB) following the attention structure given in [30]. In particular, one channel attention module and one node attention module are added after the outputs of max-pooling and average-pooling of each GCL as shown in Fig. 6.

Differently, we replace the common convolution operation with graph convolution in the second part, called node attention, to match the requirement of the GNN. Therefore, for one middle feature map H, the channel attention coefficients and the node attention coefficients are calculated as:

$$M_{ch}(H) = \odot(MLP(AvgPool(H)) + MLP(MaxPool(H))), \tag{8}$$

$$M_{no}(H') = \odot(GCL(AvgPool(H'); MaxPool(H'))), \tag{9}$$

where $\odot$ denotes *sigmoid* function, the *MLP* is a multi-layer perceptron with *ReLU* activation function that shares weights for both pooling results, the $H' = M_{ch}(H) \otimes H$, $H'' = M_{no}(H') \otimes H'$, and $\otimes$ indicates element-wise multiplication. Specifically, the pooling in channel attention step is to focus on the importance of different channels instead of node representations, while the pooling in the node attention is conducted across the channel to suppress information of redundant channels.

Note that graph attention networks (GATs) [31] also can achieve attention on graph learning. But only the edge connections are used in GATs, the consideration of the edge attributes is missing whose are very important semantic information for FER. That is why we design the GCAB instead of applying GATs directly. Based on the GCAB, a DSN can be built for feature learning from the generated facial graph.

3.2.2. Graph transformation

Before processing our facial graph by GCABs, we need firstly transform the graph data into the input format that satisfies network training. Concretely, the variant adjacency matrix $M_a \in \mathbb{R}^{N \times N}$, the node texture matrix $M_c \in \mathbb{R}^{N \times C}$ of graph and the landmark location matrix $M_l \in \mathbb{R}^{N \times 2}$ are constructed for feature learning as follows:

$$M_a = \begin{bmatrix} 0 & s_{1,2} & \cdots & \cdots & s_{1,N} \\ s_{2,1} & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & s_{N-1,N} \\ s_{N,1} & \cdots & \cdots & s_{N,N-1} & 0 \end{bmatrix}, \tag{10}$$

and

$$M_c = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1C} \\ f_{21} & f_{22} & \cdots & f_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ f_{N1} & f_{N2} & \cdots & f_{NC} \end{bmatrix}, \; M_l = \begin{bmatrix} l_{1,x} & l_{1,y} \\ l_{2,x} & l_{2,y} \\ \vdots & \vdots \\ l_{N,x} & l_{N,y} \end{bmatrix}, \tag{11}$$

where s_{pq} ($s_{pq} = s_{qp}$) is the semantic index computed depending on different strategies. And $f_{p[1,\ldots,C]}$ are local texture features calculated by Eq. (3) or extracted by lite-CNN module, while $l_{i,x}$ and $l_{i,y}$ are the landmark coordinates respectively.

3.2.3. Network architecture and learning

To capture the appearance and geometric expression patterns as well as their semantic relationships simultaneously, it is important to keep the graph structure during the learning process. Thus, we stack up several GCABs into a multi-layer dual-stream network (DSN) for multi-level expression feature learning on the facial graph. Since the local appearance and landmark positions belong to different feature spaces and have different dimensions, two streams are needed and the number of GCABs they need may also be different. Fig. 7 gives the architecture of DSN. Effects of different GCAB layer numbers are tested in Section 4.2.4 to determine the optimal network architecture.

The M_a, M_c and M_l are then fed into the DSN and the required item Z for graph convolution operation is calculated previously according to Eq. (7) before data loading. All the trainable parameters are updated by *stochastic gradient descent (SGD)* with *backprop-*

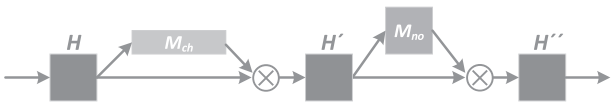


Fig. 6. The process of attention mechanism in GCAB.

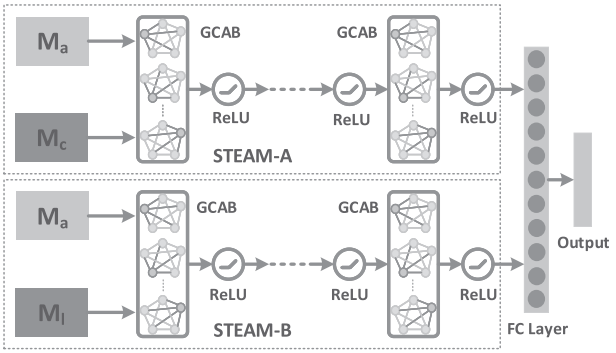


Fig. 7. The architecture of DSN.

agation (BP). After several layers of the GCAB, the outputs of each node in DSN are combined with the fully connected layer and set as the input to a *softmax* layer for expression prediction. The *cross-entropy* cost is evaluated as the loss function for model training and the *dropout* trick is applied to alleviate the overfitting problem.

4. Experiments

In this section, we evaluate the performance of SG-DSN on three lab-controlled FER datasets: Extended Cohn-Kanade (CK+) [32], Oulu-CASIA [33] and MMI [34], and two in-the-wild datasets: Static Facial Expressions in the Wild (SFEW) [35] and real-world expression database (RAF-DB) [36]. In particular, multiple groups of comparison experiments on CK + and Oulu-CASIA datasets are set up for ablation studies. Then, the recognition power of the SG-DSN model is verified against several latest methods on five datasets.

4.1. Implementation details

In this work, all parameters are trained using NVIDIA GeForce GTX 1080Ti GPU based on the open-source Tensorflow platform. Each GCAB has 64 channels for output. The decayed learning rate is set as 0.005, and the *dropout* ratio is 0.5. All the parameters are fixed throughout the whole experiments. Images from all the datasets are resized to 224×224. The landmark detection is performed by SAN [37] and the official provided metadata of landmarks are used for hard samples. For those still undetectable images, we just discard them and treat them as samples of the wrong prediction when evaluating the FER performance. In addition, the experimental environment and data preprocessing methods are the same or similar as previous approaches for a fair comparison.

4.1.1. Datasets

CK+ is one of the most widely used FER datasets, which contains 593 image sequences of 123 subjects. In this paper, seven emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise) are taken into consideration. For each sequence, the last three frames are selected and grouped into 10 subject-independent subsets for 10-fold cross-validation.

Oulu-CASIA includes six emotions (without contempt) collected from 80 subjects aged 23 to 58. Each sequence starts at the onset frame and ends at the apex frame of the corresponding expression. Similar to CK + dataset, the last three frames of every sequence are grouped for the 10-fold cross-validation.

MMI has 205 sequences of 30 subjects labeled with six emotions. In this work, we conduct a subject-independent 10-fold cross-validation for experimental comparability. It is noteworthy that three apex frames in each sequence are selected for the experiment.

SFEW consists of 1394 images from video clips of movies in the real world, which are labeled as six basic expressions, and neutral expression. Since it has divided training, validation and test groups, we select these official sets for cross-validation.

RAF-DB is a large-scale in-the-wild dataset and has 15,339 images with the same expressions as the SFEW dataset. In this work, 12,271 and 3068 images are used as the training set and test set respectively.

4.1.2. Metrics

We evaluate the proposed SG-DSN with accuracy metric to compare the performance of each FER approach quantitatively. The accuracy Acc_i of the convergence model in i^{th} fold can be obtained by:

$$Acc_i = \frac{CPL_i}{GTL_i}, \tag{12}$$

where GTL_i and CPL_i are the total number of ground truth labels and the number of correct prediction i^{th} fold separately. Then the average FER accuracy is calculated as follows:

$$Acc = \frac{1}{\eta}\sum_{i=1}^{\eta} Acc_i, \tag{13}$$

where η is the number of folds, which is varied according to different datasets in this work.

4.2. Ablation studies

The effectiveness of our proposed components in SG-DSN is examined in this subsection by FER performance experiments on CK + and Oulu-CASIA datasets.

4.2.1. Parameters of local texture descriptors

First, we evaluate the performance on different parameter sets of Gabor and HOG features. For Gabor filters, three groups of experiments are conducted with kernel sizes of $3 \times 3, 5 \times 5$, and 7×7. And these generated Gabor vectors are then averaged for later feature fusion. As shown in Table 1, the kernel size of 3×3 performs best. One possible explanation is that a smaller kernel size might result in a better description of the texture. Thus, we use 3×3 size kernel in the final implementation. For HOG descriptors, we choose the cell size of 4×4 and 8×8 with corresponding block size of 2×2 and 1×1 separately. Table 2 shows the effects on two different groups of cell sizes and block sizes of the HOG feature for FER. The comparison reveals that the highest performance occurs at 4×4 cell size and 2×2 block size, which are followed in later experiments. And the possible reason for the case of HOG is that more blocks and small cells can achieve fine-grained and small-scale gradient sampling, especially for the local patches segmented in our work.

Table 1
Comparison of different Gabor kernel size.

Kernel size	Accuracy(%)	
	CK+	Oulu-CASIA
3×3	**96.15**	**86.40**
5×5	92.86	85.23
7×7	88.11	82.68

[1]Bold value denotes best.

Table 2
Comparison of different HOG parameters.

Parameter sets		Accuracy(%)	
Cell size	Block size	CK+	Oulu-CASIA
4×4	2×2	**95.19**	**87.25**
8×8	1×1	89.10	83.63

[1]Bold value denotes best.

4.2.2. Local feature extraction methods

In Section 3.1.1, we present two methods for local feature extraction. As shown in Table 3, we verify the effectiveness of single classic features, fused features, and lite-CNN features respectively. The results demonstrate that the SG-DSN with fused features can achieve better performance than only using the other two single local features. In addition, the fused feature and the lite-CNN feature get almost equally high accuracy on CK+, while the former performs better on Oulu-CASIA. One possible explanation is that the insufficient training samples limit the effectiveness of the lite-CNN. A similar situation also applied to the case on MMI (see Table 6). But the deep features show better results on the large-scale in-the-wild dataset (see Table 7). On the other hand, it reveals that our DSN is compatible with different local features. Therefore, both of the two methods are followed in subsequent experiments.

4.2.3. Effects of edge attributes and attention

To verify the effectiveness of the two proposed strategies for the semantic relationship initialization in our facial graph, we blind the edge attributes by setting $s_j = 1, (1 \leqslant j \leqslant |E|)$. Besides, the node attributes are generated using the fused feature as above and the attention module is not used in this stage. The results in Table 4 illustrate that both the two semantic indexes can raise 7% and 5% accuracy on CK + and Oulu-CASIA respectively. Furthermore, the accuracy of the Hop distance is slightly higher than that of the Euclidean distance, which shows that the former can better represent the semantic relationship of facial changes in expressions. On the other hand, we further explore the role of our attention module. As shown in Table 4, both the two models using different semantic strategies benefits from introducing attention mechanism into DSN. These observations confirm the prior knowledge that the significant facial changes in key areas are co-occurring and have prior importance for specific expression. Therefore, we exploit these two components to make fully use of the semantic relationship encoded in the proposed facial graph.

4.2.4. Architectures of GCAB layers

In this part, we compare the performance of DSN with different architectures. Four GCAB groups are trained with the same setting respectively, using our facial graph as input. As summarized in Table 5, the best result appears at the third architecture, which has 3-layer GCAB for stream-A and 2-layer GCAB for stream-B. It is observed that the performance does not always increase with the number of GCAB layers. We believe the reason is that the dimension of our graph structure is not large, so that too deep layers will make the node features tend to converge to the same vector and gradually become indistinguishable. This phenomenon is also in line with the studies in [38,39]. Besides, we also experiment with the case of using only one stream with a concatenated matrix of the appearance and geometric attributes. Since graph convolution is a message passing and aggregation method, we implement SVM and VGG-Face as baseline models by sending averaged local textures and semantic features. From Table 5, three of the architectures outperform the baseline model and the dual-stream framework performs better than the single DSN. One possible reason is that these two types of features have different dimensions and belong to different feature spaces. The simple feature-level fusion might suppress the contribution of geometric so that we applied two branches of feature learning and decision-level fusion, which is also used in previous methods [16]. Thus, we choose the third architecture as the backbone of SG-DSN for later performance comparison with the latest methods.

Table 3
Performance with and without feature fusion.

Method	Accuracy(%)	
	CK+	Oulu-CASIA
Single Gabor feature	96.15	86.40
Single HOG feature	95.19	87.25
Fused feature	98.86	**90.88**
Lite-CNN feature	**99.23**	89.24

[1]Bold value denotes best.

Table 4
Performance with and without edge indexes and attention.

Method (Lite-CNN)	Accuracy(%)	
	CK+	Oulu-CASIA
Without edge indexes	90.81	81.59
With Euclidean distance	97.52	86.32
With Hop distance	98.07	86.53
Euclidean distance + Attention	98.36	88.79
Hop distance + Attention	99.23	89.24

Table 5
Different architectures of GCABs and DSN on CK + dataset.

Number of GCAB layers		Accuracy(%)	
Stream-A	Stream-B	CK+	Oulu-CASIA
2-layer	2-layer	94.36	84.12
2-layer	3-layer	90.10	81.48
3-layer	2-layer	**99.23**	**89.24**
4-layer	2-layer	95.82	87.15
DSN (3-layer)		95.63	74.46
DSN (2-layer)		91.24	69.28
Baseline (SVM)		87.19	63.15
Baseline (VGG-Face)		92.71	82.17

[1]Bold value denotes best.

4.3. Visualization of learned semantics

For the purpose of exploring the semantics of features, the visualization of features learned by SG-DSN is conducted. In particular, we link the normalized graph features to their related input nodes to present the semantic weights. As illustrated in Fig. 8, the larger the size and the darker the color of the node is, the greater contribution the node feature provides. For example, the nodes near the lip corner and upper lip play important roles in *happiness* and *surprise* respectively. In addition, the two additional landmarks also provide significant contributions in *contempt*, *disgust*, and *surprise*. And these observations are also consistent with the theory in physiological anatomy and cognitive neurology, which proves our SG-DSN can explicitly extract the semantic information of facial expressions and has certain interpretability.

4.4. Comparison with state-of-the-art methods

4.4.1. Performance evaluation on lab-controlled datasets

To evaluate the performance of SG-DSN with the above settings, we firstly conduct experiments on three lab-controlled datasets:

Table 6
Performances on three lab-controlled datasets.

Methods	Data	Accuracy(%)		
		CK+	Oulu-CASIA	MMI
DAUGN [18]	static image	97.67	84.28	80.11
DDL [40]	static image	*99.16*	88.26	83.67
DeRL [41]	image pair	97.30	88.00	73.23
DLP-CNN [36]	static image	95.78	/	78.46
DTAGN [42]	sequence	97.25	81.46	70.24
FER-IK [43]	static image	97.59	/	84.90
IFSL [44]	static image	98.70	/	**92.60**
MSCNN-PHRNN [16]	sequence	98.50	86.25	81.18
RCFN [17]	static image	97.94	86.94	/
SG-DSN (fused feature)	static image	98.86	**90.88**	*85.75*
SG-DSN (lite-CNN)	static image	**99.23**	*89.24*	82.64

[1] Bold values denote the best, italic values denote the second best.
[2] Fused feature and lite-CNN are two versions using different local feature extraction methods.
[3]The IFSL takes the advantage of non-deep feature in small-scale datasets so that obtains leading performance.

Table 7
Comparison on SFEW dataset.

Methods	Framework	Accuracy(%)	
		SFEW	RAF-DB
DAUGN [18]	G. + CNN	55.36	86.03
DLP-CNN [36]	CNN	51.05	84.13
IPFR [45]	deep	*57.10*	/
LDL-ALSG [46]	G. + CNN	56.50	85.53
RAN [4]	CNN + Att.	54.19	*86.90*
RCFN [17]	CNN	43.28	/
IFSL [44]	non-deep	46.50	76.90
SG-DSN (fused feature)	GNN + Att.	56.35	86.87
SG-DSN (lite-CNN)	GNN + Att.	**57.42**	**87.13**

[1] Bold values denote the best, italic values denote the second best.
[2] G. denotes using graph-based representation of facial expression and Att. indicates attention module.
[3] Fused feature and lite-CNN are two versions using different local feature extraction methods.

CK+, Oulu-CASIA, and MMI against several previous FER methods, including DAUGN [18], DDL [40], DeRL [41], DLP-CNN [36], DTAGN [42], FER-IK [43], IFSL [44], MSCNN-PHRNN [16], and RCFN [17] respectively.

CK+: From Table 6, we find that most approaches achieve the accuracy higher than 97%. Although DLP-CNN utilizes joint supervision by softmax loss and locality preserving loss, it lacks the use of spatial semantic relationships which causes lower a result. In addition, DATGN and MSCNN-PHRNN combine the appearance and geometric information and acquire good accuracy with the help of extra temporal information. By contrast, our SG-DSN gets the best in this experiment with the help of graph-based representation and the attention mechanism. Also, it performs better than DAUGN and DDL, of using the above two separately. The confusion matrix in Fig. 9(a) illustrates that our SG-DSN performs well at all the seven emotion classifications, which can be attributed to the excellent representation strategy.

Oulu-CASIA: As shown in Table 6, under the normal illumination condition of this dataset, all the compared methods get the FER accuracy over 80%. Specifically, DDL alleviates the generic knowledge of AUs by integrating a Bayesian Network into a deep learning framework and gains a remarkable performance improvement. Similar to the experiment on CK + dataset, MSCNN-PHRNN also performs well on the Oulu-CASIA dataset that further confirms the effectiveness of feature fusion technology. The top-2 results come from the proposed SG-DSN model. The fused feature beat lite-CNN because it is more capable of small-scale and low-resolution samples. The confusion matrix in Fig. 9(b) demonstrates that SG-DSN acquires satisfying results for four emotions except for *sadness* and *disgust*. One possible explanation is that samples of these two categories in Oulu-CASIA are too similar to learn discriminative features.

MMI: Different from CK + and Oulu-CASIA datasets, MMI has less image samples and more non-aligned poses. As shown in Table 6, most approaches suffer an accuracy drop with different degrees in this experiment. DTAGN and DeRL achieve FER accuracy just over 70%. The reason is the effectiveness of features extracted from these two methods highly relies on their deep architectures and sufficient training data, which is what the MMI dataset does not satisfy. In other words, that is why IFSL gets the best result by using the non-deep method. This can also explain the performance degradation of lite-CNN. Still, our SG-DSN (fused feature) integrates local-to–holistic expression information based on semantic relationships and acquires the second-best FER accuracy. From the confusion matrix presented in Fig. 9(c), we summarize that SG-DSN performs well at *disgust*, *happiness,* and *surprise*. The poor results appear in *anger*, *fear,* and *sadness*, which can be imputed to the insufficient and unbalanced training data of corresponding expressions.

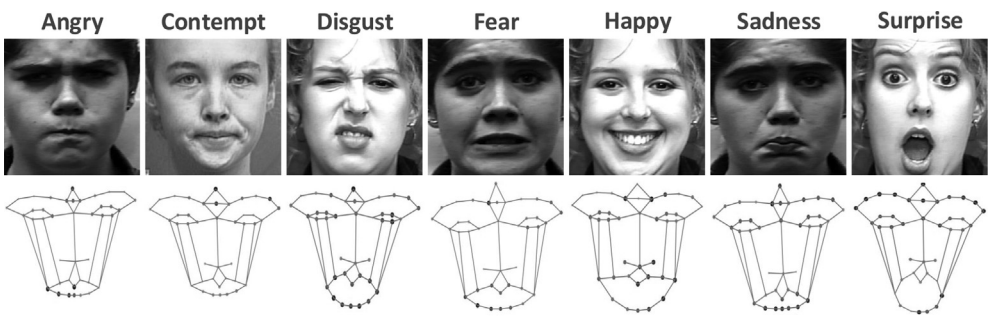
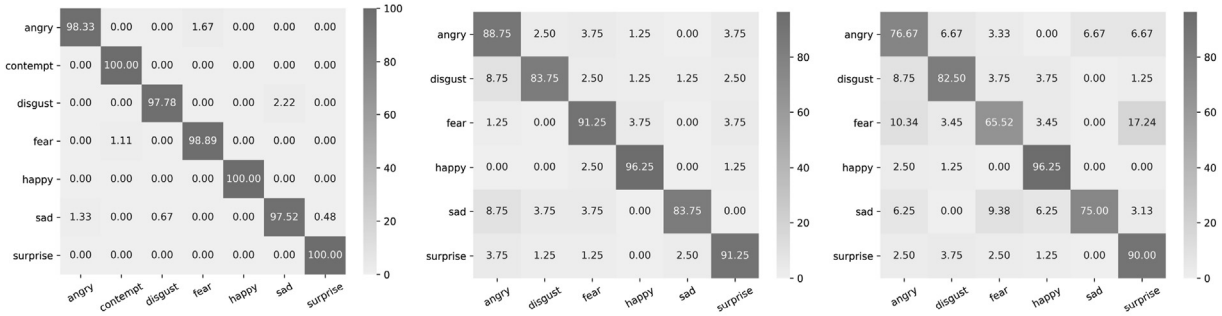


Fig. 8. The visualization of the learned semantic features of SG-DSN.



(a) CK+ (b) Oulu-CASIA (c) MMI

Fig. 9. The confusion matrices on lab-controlled datasets.

4.4.2. Performance evaluation on in-the-wild datasets

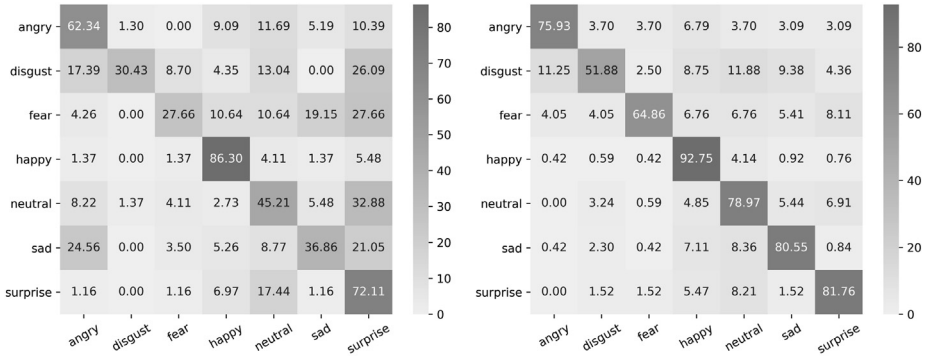
To further verify the effectiveness of our SG-DSN in real scenarios, we conduct the experiment on SFEW and RAF-DB datasets and add three state-of-the-art comparison methods RAN [4], IPFR [45], and LDL-ALSG [46].

SFEW: As shown in Table 7, due to the challenging factors, RCFN fails to break through the accuracy over 50%. This is because its region segmentation is heavily weakened in real environments. LDL-ALSG performs well by utilizing a label space graph that links multiple labels with different intensities to one emotion. RAN adaptively captures the importance of facial regions to handle the pose variant, and achieves an accuracy of 54.19%. Our SG-DSN defeats other methods, even if the face alignment and landmark coordinates are not precise enough. This shows that the semantic relationship can also enhance recognition performance in real scenes. The confusion matrix in Fig. 10(a) illustrates that SG-DSN can recognize *anger*, *happiness,* and *surprise* well, but barely distinguish *disgust,* and *fear*. The main reason is that the facial deformation and background interference in the SFEW dataset impair the power of our facial graph.

RAF-DB: According to Table 7, DAUGN also performs well by using graph-based representation. Although the RAF-DB dataset provides sufficient samples, the non-deep framework of the IFSL makes it difficult to benefit from such an advantage, so that causes a big gap of accuracy. The proposed SG-DSN using lite-CNN obtains the best result, we think this is due to our effective facial graph and the powerful DSN. Because of the limitation of hand-crafted features in such a large-scale in-the-wild dataset, our fused features suffer an accuracy decrease, but it still achieves a favorable performance just slightly lower than RAN. The confusion matrix in Fig. 10 (b) indicates that SG-DSN performs well at all the emotion categories except *disgust* and *fear*, which can be attributed to the large training data in RAF-DB.

4.5. Discussion

Although the five datasets have different conditions, the proposed SG-DSN outperforms most of the compared state-of-the-art approaches. In CK + dataset, the high quality of images makes it possible to conduct precise facial landmark detection that ensures the validity of the graph representation and achieve remarkable FER results. In the Oulu-CASIA dataset, for the sake of the graph convolution process, our SG-DSN overcomes the low-quality problem and extracts powerful expression features. And in the MMI dataset, we notice that the FER accuracy is influenced by the not-well-aligned faces which decrease the useful semantic information of the generated graph. Alternatively, our SG-DSN still gets competitive performance due to the effective local features and the combination of appearance and geometric information. When facing the in-the-wild SFEW and RAF-DB datasets, the proposed SG-DSN keeps an impressive accuracy in the case of the large facial deformation and the complex background. Nevertheless, this also reveals that our graph representation is closely related to the accuracy of landmark detection, which still limits the performance of SG-DSN in real-world environments. For the two local feature extraction modules, we experimentally find that the fused classic feature is better at handling small-scale posed facial data, while the lite-CNN is more flexible for large-scale real data.



(a) SFEW (b) RAF-DB

Fig. 10. The confusion matrices on in-the-wild datasets.

5. Conclusion and future work

In this paper, we presented a novel model for FER, the Semantic Graph-based Dual-stream Network (SG-DSN). This method generated a facial graph to represent significant facial changes and their semantic relationships. In addition, SG-DSN exploited a network with graph convolutional attention blocks (GCABs) to jointly learn powerful features from both appearance and geometric expression information. The proposed edge indexing initialization strategies made full use of the semantic relationship based on prior knowledge and improved the performance in expression recognition. On both lab-controlled and in-the-wild challenging datasets, SG-DSN achieved competitive FER results. In the future, a facial graph representation without relying on landmarks can be explored to account for more complex scenarios and the dynamic evolution of facial expressions can be considered.

CRediT authorship contribution statement

Yang Liu: Conceptualization, Methodology, Software, Writing - original draft. **Xingming Zhang:** Project administration, Writing - review & editing. **Jinzhao Zhou:** Conceptualization, Data curation, Visualization. **Lunkai Fu:** Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding: This work was supported by the China Scholarship Council (CSC, No. 202006150091).

Appendix A. Ablation Study: Effects on different numbers of landmarks

Since landmark selection and addition are employed in this work, we further conduct an experiment to evaluate its contribution. To fair comparison, we use the Euclidean distance here because the hop distance of each graph node is indistinguishable. As shown in Table A.8, the FER performance on both posed and in-the-wild datasets obtains a significant increase after using the landmark selection. One possible explanation is that the fully connected graph based on all the detected landmarks contains unnecessary nodes and edges, which distracts the importance of crucial graph nodes, especially when dealing with uncontrolled faces. And the experimental results also reveal that the additional nodes and accompanying edges provide important spatial and semantic information.

Table A.8
Performance with or without landmark selection.

Method (lite-CNN)	Accuracy(%)	
	CK+	SFEW
Without selection	95.48	45.70
Without addition	97.56	50.55
With both	**98.36**	**53.14**

[1] Bold value denotes best.

References

[1] I. Gogić, J. Ahlberg, I.S. Pandžić, Regression-based methods for face alignment: a survey, Signal Processing 178 (2021) 107755.
[2] Z. Fei, E. Yang, D.D.-U. Li, S. Butler, W. Ijomah, X. Li, H. Zhou, Deep convolution network based emotion analysis towards mental health care, Neurocomputing 388 (2020) 212–227.
[3] A.R. Kurup, M. Ajith, M.M. Ramón, Semi-supervised facial expression recognition using reduced spatial features and deep belief networks, Neurocomputing 367 (2019) 188–197.
[4] K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, IEEE Transactions on Image Processing 29 (2020) 4057–4069.
[5] J. Lee, S. Kim, S. Kim, J. Park, K. Sohn, Context-aware emotion recognition networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 10143–10152.
[6] J. Shao, Y. Qian, Three convolutional neural network models for facial expression recognition in the wild, Neurocomputing 355 (2019) 82–92.
[7] S. Xie, H. Hu, Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks, IEEE Transactions on Multimedia 21 (1) (2018) 211–220.
[8] D.H. Kim, W.J. Baddar, J. Jang, Y.M. Ro, Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition, IEEE Transactions on Affective Computing 10 (2) (2017) 223–236.
[9] M.G. Calvo, A. Gutiérrez-García, M. Del Líbano, What makes a smiling face look happy? visual saliency, distinctiveness, and affect, Psychological Research 82 (2) (2018) 296–309.
[10] R. Ekman, What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), Oxford University Press, USA, 1997.
[11] D.L. Bimler, G.V. Paramei, Facial-expression affective attributes and their configural correlates: components and categories, Spanish Journal of Psychology 9 (1) (2006) 19.
[12] S. Mohseni, N. Zarei, S. Ramazani, Facial expression recognition using anatomy based facial graph, in: Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2014, pp. 3715–3719..
[13] A. Dapogny, K. Bailly, S. Dubuisson, Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection, International Journal of Computer Vision 126 (2–4) (2018) 255–271.
[14] K. Wen, X. Gu, Q. Cheng, Learning dual semantic relations with graph attention for image-text matching, IEEE Transactions on Circuits and Systems for Video Technology doi:10.1109/TCSVT.2020.3030656..
[15] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S. Y. Philip, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems doi:10.1109/TNNLS.2020.2978386..
[16] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Transactions on Image Processing 26 (9) (2017) 4193–4203.
[17] Y. Ye, X. Zhang, Y. Lin, H. Wang, Facial expression recognition via region-based convolutional fusion network, Journal of Visual Communication and Image Representation 62 (2019) 1–11.
[18] Y. Liu, X. Zhang, Y. Lin, H. Wang, Facial expression recognition via deep action units graph network based on psychological mechanism, IEEE Transactions on Cognitive and Developmental Systems 12 (2) (2020) 311–322.
[19] G. Li, X. Zhu, Y. Zeng, Q. Wang, L. Lin, Semantic relationships guided representation learning for facial action unit recognition, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI), 2019, pp. 8594–8601.
[20] L. Zhong, C. Bai, J. Li, T. Chen, S. Li, Y. Liu, A graph-structured representation with brnn for static-based facial expression recognition, in: Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG), IEEE, 2019, pp. 1–5.
[21] M. Zhang, Y. Liang, H. Ma, Context-aware affective graph reasoning for emotion recognition, in: Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), 2019, pp. 151–156.
[22] F.P. Such, S. Sah, M.A. Dominguez, S. Pillai, C. Zhang, A. Michael, N.D. Cahill, R. Ptucha, Robust spatial filtering with graph convolutional neural networks, IEEE Journal of Selected Topics in Signal Processing 11 (6) (2017) 884–896.
[23] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, M. Pantic, Deep structured learning for facial expression intensity estimation, Image and Vision Computing 259 (2017) 143–154.
[24] A. Gudi, H. E. Tasli, T. M. Den Uyl, A. Maroulis, Deep learning based facs action unit occurrence and intensity estimation, in: Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6, IEEE, 2015, pp. 1–5..
[25] J. Zhou, X. Zhang, Y. Liu, X. Lan, Facial expression recognition using spatial-temporal semantic graph network, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020, pp. 1961–1965.
[26] Y. Li, X. Huang, G. Zhao, Micro-expression action unit detection with spatial and channel attention, Neurocomputing 436 (2021) 221–231.
[27] M.S. Gazzaniga, R.B. Ivry, G. Mangun, Cognitive Neuroscience. The Biology of the Mind, 4th ed., 2013.
[28] E. Friesen, P. Ekman, Facial action coding system: a technique for the measurement of facial movement, Palo Alto 3..

[29] O. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: British Machine Vision Association, 2015, pp. 1–12, https://ora.ox.ac.uk/objects/uuid:a5f2e93f-2768-45bb-8508-74747f85cad1.
[30] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19..
[31] P. Veliković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: International Conference on Learning Representations, 2018, pp. 1–12. https://openreview.net/forum?id=rJXMpikCZ..
[32] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: Proceedings of the 2010 IEEE Computer Society Sonference on Computer Vision and Pattern Recognition-workshops (CVPR), IEEE, 2010, pp. 94–101.
[33] G. Zhao, X. Huang, M. Taini, S.Z. Li, M. PietikäInen, Facial expression recognition from near-infrared videos, Image and Vision Computing 29 (9) (2011) 607–619.
[34] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the mmi facial expression database, in: Proceedings of the 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Paris, France, 2010, p. 65..
[35] A. Dhall, R. Goecke, J. Joshi, K. Sikka, T. Gedeon, Emotion recognition in the wild challenge 2014: Baseline, data and protocol, in: Proceedings of the 16th International Conference on Multimodal Interaction (ICMI), 2014, pp. 461–466.
[36] S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Transactions on Image Processing 28 (1) (2019) 356–370.
[37] X. Dong, Y. Yan, W. Ouyang, Y. Yang, Style aggregated network for facial landmark detection, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 379–388.
[38] G. Li, M. Muller, A. Thabet, B. Ghanem, Deepgcns: Can gcns go as deep as cnns?, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 9267–9276..
[39] K. Zhou, Y. Dong, W. S. Lee, B. Hooi, H. Xu, J. Feng, Effective training strategies for deep graph neural networks, arXiv preprint arXiv:2006.07107. https://arxiv.org/abs/2006.07107..
[40] D. Ruan, Y. Yan, S. Chen, J.-H. Xue, H. Wang, Deep disturbance-disentangled learning for facial expression recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2833–2841.
[41] H. Yang, U. Ciftci, L. Yin, Facial expression recognition by de-expression residue learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2168–2177.
[42] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2983–2991.
[43] Z. Cui, T. Song, Y. Wang, Q. Ji, Knowledge augmented deep neural networks for joint facial expression and action unit recognition, in: Advances in Neural Information Processing Systems (NIPS), vol. 33, 2020, pp. 14338–14349..
[44] Y. Yan, Z. Zhang, S. Chen, H. Wang, Low-resolution facial expression recognition: A filter learning perspective, Signal Processing 169 (2020) 107370.
[45] C. Wang, S. Wang, G. Liang, Identity-and pose-robust facial expression recognition through adversarial feature learning, in: Proceedings of the 27th ACM International Conference on Multimedia (MM), 2019, pp. 238–246.
[46] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, Y. Rui, Label distribution learning on auxiliary label space graphs for facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13984–13993.

YANG LIU received the B.Eng. and M.Eng. degree in computer science and technology from the Nanjing Normal University and Guangxi University, China, in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree in computer science and technology from South China University of Technology, Guangzhou, China. Since 2020, he is a visiting scholar in the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. His research interests include facial expression recognition, affective computing and machine learning.

XINGMING ZHANG received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1996. He is currently a Professor, a Doctoral Supervisor, and the Vice Dean with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He is also the Executive Director of the Computer Federation of Guangdong Province. His research interests focus on data analysis, image processing, video coding, and surveillance. He is a Standing Committee Member of the Technical Committee of Education, China Computer Federation.

JINZHAO ZHOU received the B.Eng. degree from the School of Economics and Commerce, South China University of Technology, Guangzhou, China, in 2018, where he is currently pursuing the M.S. degree from the School of Computer Science and Engineering. His research interests focus on the field of image processing and facial expression recognition with deep learning.

LUNKAI FU received B.Eng. degree in computer science from the Guangdong Polytechnic Normal University, China, in 2019. He is currently pursuing the M.S. degree in computer science and technology from South China University of Technology, Guangzhou, China. His research interests include facial expression recognition, weather recognition and computer vision.