Pattern Recognition 103 (2020) 107321

Contents lists available at ScienceDirect

Pattern Recognition

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

Graph convolutional network with structure pooling and joint-wise channel attention for action recognition

Yuxin Chen[a,1], Gaoqun Ma[b,1], Chunfeng Yuan[a,*], Bing Li[a,*], Hui Zhang[e], Fangshi Wang[b], Weiming Hu[a,c,d]

[a] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China
[b] School of Software Engineering, Beijing Jiaotong University, Beijing 100044, PR China
[c] CAS Center for Excellence in Brain Science and Intelligence Technology Academy of Sciences, Beijing 100190, PR China
[d] University of Chinese Academy of Sciences, Beijing 100190, PR China
[e] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, PR China

ARTICLE INFO

Article history:
Received 1 June 2019
Revised 22 January 2020
Accepted 28 February 2020
Available online 29 February 2020

MSC:
00-01
99-00

Keywords:
Graph convolutional network
Structure graph pooling
Joint-wise channel attention

ABSTRACT

Recently, graph convolutional networks (GCNs) have achieved state-of-the-art results for skeleton based action recognition by expanding convolutional neural networks (CNNs) to graphs. However, due to the lack of effective feature aggregation method, e.g. max pooling in CNN, existing GCN-based methods only learn local information among adjacent joints and are hard to obtain high-level interaction features, such as interactions between five parts of human body. Moreover, subtle differences of confusing actions often hide in specific channels of key joints' features, this kind of discriminative information is rarely exploited in previous methods. In this paper, we propose a novel graph convolutional network with structure based graph pooling (SGP) scheme and joint-wise channel attention (JCA) modules. The SGP scheme pools the human skeleton graph according to the prior knowledge of human body's typology. This pooling scheme not only leads to more global representations but also reduces the amount of parameters and computation cost. The JCA module learns to selectively focus on discriminative joints of skeleton and pays different levels of attention to different channels. This novel attention mechanism enhance the model's ability to classify confusing actions.

We evaluate our SGP scheme and JCA module on three most challenging skeleton based action recognition datasets: NTU-RGB+D, Kinetics-M, and SYSU-3D. Our method outperforms the state-of-art methods on three benchmarks.

1. Introduction

Human action recognition is an active research area in computer vision, since it plays a critical role in video understanding and has important applications in many areas, such as video surveillance, human-machine interaction, and virtual reality [1–3]. Human action can be recognized from multiple modalities of data, such as RGB videos, depth sequences, skeleton data, etc. Among these modalities, the amount of skeleton data is growing rapidly [4,5] due to high-precision depth sensors like Microsoft Kinect v2 and advanced pose estimation algorithms [6]. Different from RGB videos, skeleton data only includes 3D locations of major body joints [7]. Such representation is robust to variation of backgrounds and viewpoints. Thanks to its high level representation and robustness, action recognition based on skeleton data [8–13] has attracted increasing attention.

Most of current methods adopt either recurrent neural networks (RNNs) [14–16] or convolutional neural networks (CNNs) [17–21] for action recognition. They treat the skeleton sequences as time series or pseudo-images and feed them into RNNs or CNNs to extract human motion features. However, the skeleton of human body is naturally a graph structure, not a sequence or a rigid image. These methods can not capture the spatial relationships between the joints, which are crucial for understanding human actions.

* Corresponding authors.
E-mail addresses: chenyuxin2019@ia.ac.cn (Y. Chen), 17121711@bjtu.edu.cn (G. Ma), cfyuan@nlpr.ia.ac.cn (C. Yuan), bli@nlpr.ia.ac.cn (B. Li), zhanghui@iie.ac.cn (H. Zhang), fshwang@bjtu.edu.cn (F. Wang), wmhu@nlpr.ia.ac.cn (W. Hu).
[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.patcog.2020.107321
0031-3203/

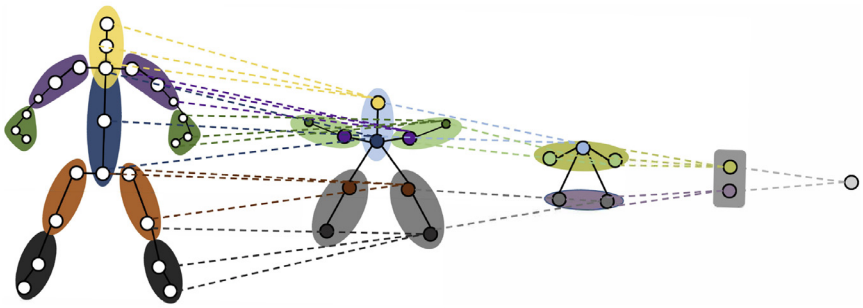


Fig. 1. The Structure based graph pooling scheme. According to the inherent structure and movement characteristics of the human body, gradually expand the receptive field and learn high-lever features.

Graph neural networks have been explored and made advance in a wide range of research fields. Typically, [22,23] utilized GNN to pass the interaction message between human and human or human and objects. Wang et al. [24] constructed a graph according to the proposed grammars. In [25,26], the human body was represented as a graph and GNN was utilized to incorporate structural knowledge. In [27], a novel GNN was proposed to capture higher-order relations between video frames. Zheng et al. [28] formalized the visual dialogs task as inference in a graphical model with partially observed nodes and unknown graph structures. Some methods [29–31] apply Graph Convolutional Networks (GCNs) to skeleton based action recognition to exploit the spatial relationships between the joints by generalizing convolutions to graph domain. Although the GCNs have achieved good results in skeleton based action recognition, few of them explored how to define pooling operation for human skeleton, which is significant for extracting global motion features. In this paper, we propose a structure based graph pooling scheme for human skeleton. Compared with previous graph pooling methods [32,33], the proposed SGP scheme introduces little computation and is easy to train. It contains prior knowledge of human skeleton typology and thus is suitable for skeleton based action recognition. The SGP scheme gradually pools the human skeleton graph and expands receptive fields of graph convolution kernel without destroying topological structure of human skeleton. Specifically, human body can be decomposed into five parts, two arms, two legs, and one trunk. Furthermore, it can be roughly divided into upper body and lower body. Using SGP scheme, our model will iteratively learn hierarchical representations of human skeleton (Fig. 1). Moreover, the SGP scheme brings a reduction on the number of parameters and computation cost significantly, which makes our model lightweight.

This work also proposes a novel joint-wise channel attention(JCA) module to distinguish confusing actions. [15,34] assign an attention weight to each joint, they pay no attention to different channels of each joint. However, subtle differences of confusing actions often hide in specific channels of key joints. For example, “reading” and “writing” are two actions that the current models are easy to confuse. The key joints of these actions are fingers. Trajectories of different channels of finger is illustrated in Fig. 2. Only the first channel is discriminative enough to distinguish two actions. The proposed JCA module applys different attention to the different channels of each joint. The information hidden in discriminative channels of key joints can be enhanced by our JCA module and the joints or channels containing redundant information are suppressed. Therefore, our method can effectively extract local nuances features and classify confusing actions.

Experiments demonstrate that the accuracies of the proposed model achieves 6.2%(cross-subject evaluation metric) and 4.9%(cross-view evaluation metric) improvement compared with the baseline model on NTU-RGB+D dataset, respectively. It outperforms previous state-of-art methods.

The main contributions of this work are summarized as follows:

- We propose a novel structure based graph pooling (SGP) scheme to gradually pool the human skeleton graph and expand the receptive fields of graph convolution kernels in deeper layers, which can enhance the ability of GCNs for extracting more global motion information and bring a reduction on the amount of parameters and computation cost.
- We propose a joint-wise channel attention (JCA) module to mine discriminative information among confusing actions with attention mechanism, which shows significant improvement for classifying confusion actions.
- Experimental results demonstrate that our method outperforms existing state-of-the-art methods.

2. Related work

With wide applications such as intelligent surveillance, human-computer interaction, etc., action recognition becomes an active research topic and draws more attention from researchers. In recent decades, many methods have been proposed to analyze human actions. Compared with traditional methods, deep learning methods have been widely used for its good performance on modeling skeleton sequences in this field. In this section, we briefly review the development of skeleton based action recognition.

2.1. Traditional methods for skeleton based action recognition

Traditional methods design handcrafted features to capture the dynamics of joint motion and represent actions. Vemulapalli et al. [11] utilized Lie group to recognize actions by taking the skeletal translations and rotations as input. Weng et al. [13] proposed Spatial-Temporal-NBNN, an extension of Naive-Bayes Nearest-Neighbor (NBNN) [35], which applied stage-to-class distance to skeleton based action recognition. Hu et al. [36] designed heterogeneous features for rgb-d activity recognition. Koniusz et al. [37] presented two kernel-tensor to capture the spatio-temporal compatibility of body-joints and dynamics of each sequence explicitly. Wang et al. [38] proposed a novel graph kernel method to measure the similarities between skeletal human actions. However, most of these methods extract handcrafted features and then use SVM for classification, which is difficult to adapt large datasets.

2.2. CNNs and RNNs based methods for skeleton based action recognition

In recent years, there is a great success of deep learning methods in computer vision tasks, many models are proposed for skeleton based action recognition. These works can be divided into two categories: CNNs based models, and RNNs based models. As for CNNs based models, Ke et al. [17] presented a method for spatial

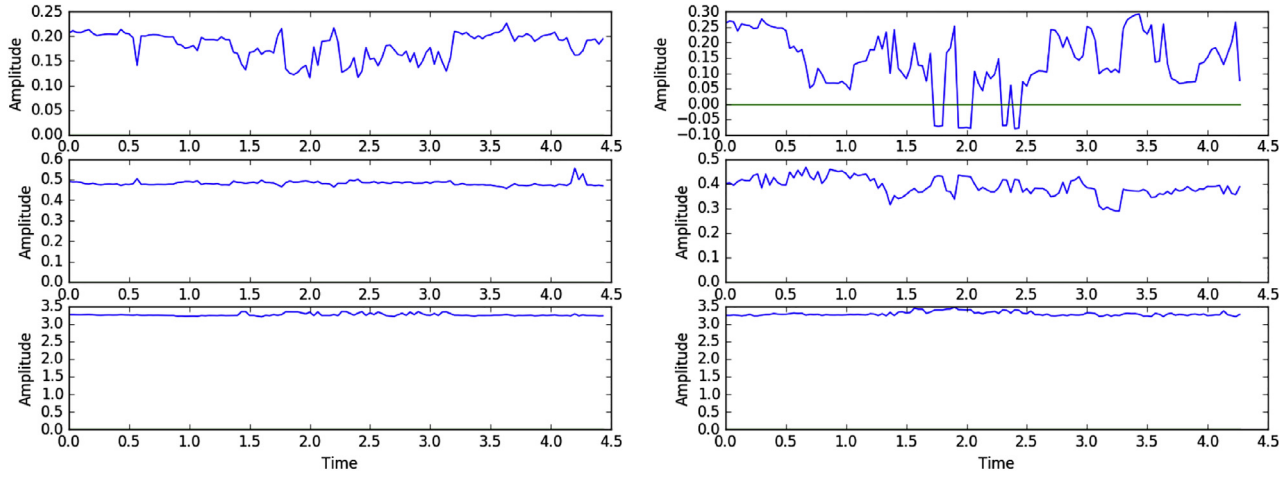


Fig. 2. Visualization of three channels for "reading" and "writing". The left column is "reading", and the right column is "writing".

temporal learning using CNNs by transforming each skeleton sequence into three clips each consisting of several frames. Liu et al. [18] developed a view-independent method to eliminate the effect of viewpoint variations, described skeleton sequences as color images, and served them to CNNs for classification. However, CNNs based methods can only capture the information in the spatial domain, ignoring most of information in the temporal domain. As for RNNs based models, which are effective in capturing temporal information, are widely applied in skeleton based action recognition. Du et al. [40] introduced an end-to-end hierarchical RNN, which divided the raw skeletons into five parts and fed each part into the RNNs. Zhu et al. [14] claimed that the inherent co-occurrences of skeletons were neglected, and thus proposed a new method with Long Short-Term Memory (LSTM) to capture the co-occurrence information. Shahroudy et al. [42] proposed a part-based LSTM to extract features for each part of human body respectively. Liu et al. [16] introduced a new gating mechanism within LSTM to analyze the reliability of input data. Song et al. [15] proposed a spatial and temporal attention model for skeleton based action recognition. However, RNN and LSTM are computationally consuming model. These models cannot be further deepened, resulting in poor representations in the spatial domain.

2.3. GCNs based methods for skeleton based action recognition

The CNNs based and RNNs based models usually ignore the inherent structure of the human body, while GCNs based methods can extract structural features by convolution in the neighborhood of joints. As the skeleton of human body is naturally a graph structure, applying graph convolution to skeleton based action recognition is intuitively very sensible. Duvenaud et al. [44], Niepert et al. [45], Yan et al. [29] proposed spatial temporal graph convolution networks(ST-GCN) to capture the patterns embedded in the joints and proved GCNs can learn better skeleton based action representations. An attention mask was employed to emphasize the effect of key joints. But it's still hard to distinguish confusing actions without going deeper into channel dimension. Tang et al. [30] presented a deep progressive reinforcement learning method based on GCNs, which selected the most informative skeletal frames and discarded ambiguous frames.Wen et al. [31] utilized motif-based graph convolution to encode hierarchical spatial structure of skeleton and a variable temporal dense block to exploit local temporal information over different ranges of human skeleton sequences. However, these works lack pooling operations between graph convolutional layers. This limits the receptive field scales of the GCNs and the model's ability to capture global motion features.

3. Proposed method

In this section, we first illustrate and outline the architecture of the proposed GCN with SGP and JCA. Then we introduce the graph construction and graph convolution in the graph convolutional layer. Next, we define the proposed SGP and introduce the overall pooling scheme in detailed. At last, we present the details of our joint-wise channel attention module.

3.1. Pipeline overview

We propose a novel graph convolutional network with SGP scheme and JCA modules. The pipeline of our model is shown in Fig. 3. We take the skeleton with 25 joints for example. For a skeleton sequences, we first extract local features between adjacent joints using a graph convolutional layer. Then we pools the original skeleton graph with 25 nodes into a new graph with 10 nodes using the 1st SGP operation. To extract discriminative features hidden in the specific channels, we employ a JCA module after the 1st SGP operation. In the same way, we consecutively connect multiple composite layers of this kind, where each composite layer is composed of a GC, a SGP and a JCA. In this paper,

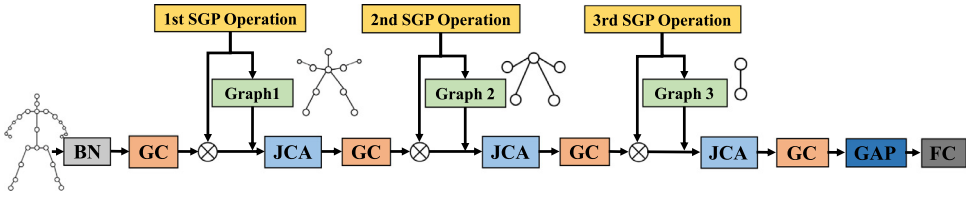
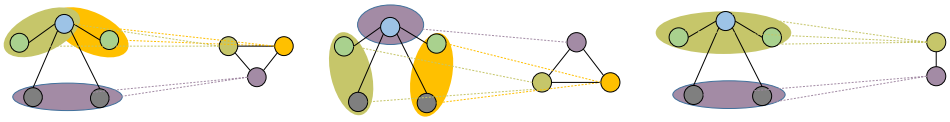


Fig. 3. The architecture of the proposed method for skeleton based action recognition. It contains four graph convolutional(GC) layers, three SGP operations and three JCA modules connected alternatively. Each SGP operation employs a transformation matrix for pooling and generates a graph for next GC layers. ⊗ denotes matrix multiplication.



(a) The first approach (b) The second approach (c) The third approach

Fig. 4. Three division approaches of the 3rd SGP operation.

we use three composite layers and one single GC layer to extract high-level interaction features. For temporal features extraction, we employ a temporal convolutional layer and a temporal Max Pooling with stride 2 after each GC layer. To obtain the final feature vector F, the global average pooling is employed to aggregate the features in spatial-temporal domain. At last, the final feature vector F is fed into the fully-connected layer to produce classification scores. Next, we firstly introduce the graph construction and convolution in each single GC layer.

3.2. Graph convolutional layer

Graph Construction. The human body can be naturally considered as an articulated system consisting of hinged joints and rigid bones, which inherently lies in a graph-based structure. Thus, we construct a graph $G = \{V, E\}$ to model the human body in each single frame, where V is the set of n joint nodes in graph and E is the set of edges. Let $X \in R^{n \times m}$ denote the joint feature matrix assuming that each joint has a m-dimensional feature and $A \in \{0, 1\}^{n \times n}$ denote the adjacency matrix, where A_{ij} is defined as:

$$A_{ij} = \begin{cases} 1 & \text{if i=j or joint i and j are connected,} \\ 0 & \text{if joint i and joint j are disconnected.} \end{cases} \tag{1}$$

Graph Convolution. Based on the constructed graph, the graph convolution response of joint v_i is computed as:

$$Y(v_i) = \sum_{v_j \in N(v_i)} \frac{1}{Z_i(v_j)} X(v_j) W(\ell(v_j)), \tag{2}$$

where $X(v_j)$ is the feature of joint v_j, and $Y(v_i)$ is the response of graph convolution operation at joint v_j. The neighbor set of the joint v_i is defined as $N(v_i) = \{v_j | d(v_i, v_j) \le D\}$, where $d(v_i, v_j)$ is the minimum distance between v_i and v_j. $Z_i(v_j)$ is the number of joints in the corresponding neighbor set, which normalizes the feature representations. A joint labeling function ℓ: $v_i \to \{1, 2, ..., K\}$ is designed to assign the labels $\{1, 2, ..., K\}$ to each joint v_i, which aims to partition the neighbor set $N(v_i)$ of v_i into a fixed number of K subsets. $W(.)$ is a function that maps trainable weights to each partition group.

In each frame, the response of graph convolution on the whole graph of all joints can be represented as:

$$Y = \sum_{k=1}^{K} D_k^{-\frac{1}{2}} A_k D_k^{-\frac{1}{2}} X W_k, \tag{3}$$

where $D_k \in R^{n \times n}$ is a diagonal degree matrix and $D_k^{i,i} = \sum_j A_k^{i,j}$. A_k is the adjacency matrix corresponding to K subsets. Note that $\sum_{k=1}^{K} A_k = A$.

3.3. Structure based graph pooling

Pooling operation on graph is complicated due to graph's complex topology structure. However, the human skeleton, being a highly regular graph, can be pooled simply and effectively. Without loss of generality, we take skeleton in the NTU-RGB+D dataset for example to show how to pool the joints of a human body reasonably by our strategy. Note that our pooling strategy can be extended to any human skeleton data.

Suppose that human body is decomposed into n parts. Namely we divide the skeleton graph G with l nodes into n subgraphs g_i, $i \in \{1, 2, ..., n\}$ corresponding to n parts of human body and an edge set e_o connecting subgraphs. Each subgraph g_i contains a set of joints J_i. According to the division of subgraphs, we construct an assign matrix $P \in R^{n \times l}$, where the $P_{i,j}$ is defined as:

$$P_{ij} = \begin{cases} 1 & \text{if joint j} \in J_i, \\ 0 & \text{if joint j} \notin J_i. \end{cases} \tag{4}$$

Then the SGP operation can be formulated as:

$$Y_{out} = M \odot \widetilde{P} X, \tag{5}$$

where $Y_{out} \in R^{n \times m}$ is the output of SGP operation, $M \in R^{n \times l}$ is a trainable mask which adjust the contribution of each joint to Y_{out}, $\widetilde{P}$ is the normalized assign matrix of P, and $\odot$ denotes the Hadamard product. After this SGP operation, we obtain a new graph which consists of n new nodes corresponding to the n subgraphs and the edge set e_o.

3.4. Pooling scheme

The proposed hierarchical SGP scheme consists of three SGP operations which are defined as Eq. (5).

The 1st SGP operation pools 25 nodes into 10 nodes. This operation reduces the redundant information and accelerates the computation speed in the later layers. The 2nd SGP operation pools 10 nodes into 5 nodes. The 5 nodes obtained from this layer correspond to the five parts of human body including four limbs and trunk. The motion information contained in human body is highly related with the interaction of five body parts. Thus, graph convolution on the graph with these 5 nodes can capture more global motion information.

We design three division approaches of subgraphs for the 3rd SGP operation. Each approach represents a deep understanding of human action. The first approach divides the 5-nodes graph into three subgraphs. As illustrated in Fig. 4(a), the trunk-node (blue node) and left-hand-node are in a subgraph, the trunk-node and right-hand-node are in a subgraph, the leg-nodes are in another subgraph. This approach considers the motion of human body consisting of interactions among left-upper body, right-upper body and legs. The second approach shown in Fig. 4(b) also divides the original graph into three subgraphs. The left-hand-node and left-leg-node belongs to one subgraph. The right-hand-node and right-leg-node belongs to one subgraph. The trunk-node is in another subgraph. It divides the whole body motion into the motion of left-half body, right-half body and the trunk. Fig. 4(c) shows the third approach, which divides the graph into two nodes corresponding to upper body and lower body. This approach fuses the motion features of upper body and lower body respectively and generate a new graph containing two nodes. Convolving on the generated graph can extract the interactions between upper body and lower body.

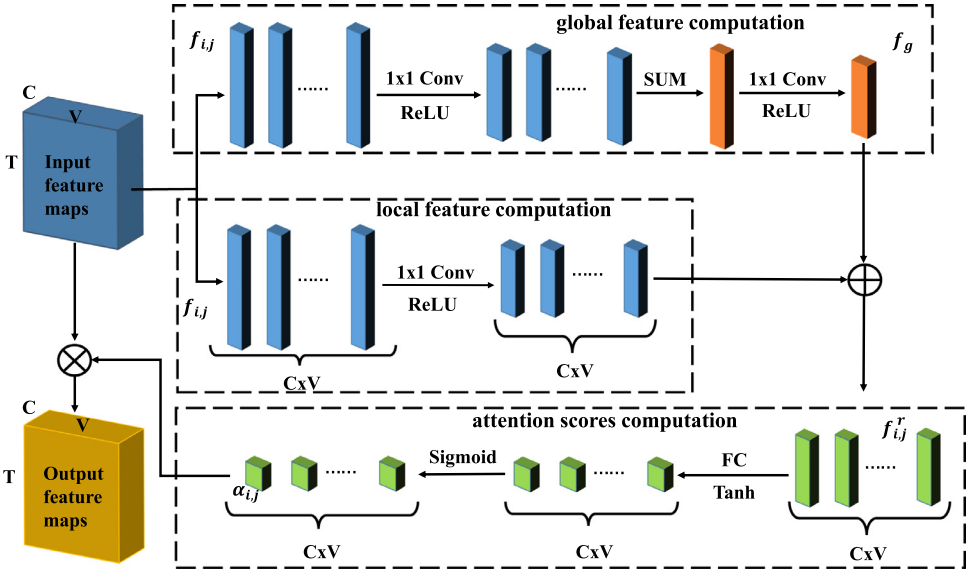


Fig. 5. Detailed structure of the proposed joint-wise channel attention (JCA) module. After JCA, every channel of each joint will get an attention weight.

3.5. Joint-wise channel attention module

The joint-wise channel attention module is proposed to focus on key channels of key joints. The detailed structure of the proposed JCA module is shown in Fig. 5. Given 3D feature maps, two branches are connected to generate a global feature and local features respectively. Then the local features plus the global feature are used to compute the final joint-wise channel attention score.

Specifically, the input of JCA module is a tensor $S \in R^{C \times T \times V}$, where C, T, V are the number of channels, frames and joints respectively. Denote the temporal vector $S[i, :, j]$ in the ith channel and jth joint of S as $f_{i,j}$. Note that if S is original data, $f_{i,j}$ is actually the trajectory of joint j on dimension i. Otherwise, $f_{i,j}$ is a specific motion pattern of local area around joint j. In one branch of the JCA module, we first use 1×1 convolution as temporal filter to extract intermediate features from each $f_{i,j}$. These intermediate features are summed up and applied with another 1×1 convolution to obtain a global feature f_g. f_g can be formulated as:

$$f_g = ReLU(W_2(\sum_{i=1}^{C}\sum_{j=1}^{|V|} ReLU(W_1 f_{i,j} + b_1)) + b_2), \tag{6}$$

where W_1, b_1, W_2, b_2 are the learnable parameters. f_g represents the motion pattern of the whole body. Then we use the other branch to extract intermediate features with same process. The intermediate features are high-level representations of their corresponding $f_{i,j}$. Each local intermediate feature is summed with f_g to obtain a group of relative motion features f_r. f_r represents the relationship between local motion pattern and global motion pattern.

At last, a fully-connected layer is employed to convert f_r to attention scores. This fully-connected layer learns to find the key channels according to f_r, i.e. the relationship between local motion pattern and global motion pattern. The attention scores are mapped into 0-1 with Sigmoid function. The final attention weight of ith channel of jth joint can be calculated as:

$$\alpha_{i,j} = Sigmoid(tanh(W_r(ReLU(W_l f_{i,j} + b_l) + f_g) + b_r)), \tag{7}$$

where W_r, b_r, W_l, b_l are the learnable parameters. We use Sigmoid as activation function for the existence of multiple key channels. Note that joint-wise attention is a special case of our joint-wise channel attention where attention scores of all channels of a joint $\alpha_{:,j}$ have the same value. Finally, the attention weights will be multiplied with its corresponding joints and channels in each frame to focus on the discriminative joints and channels.

4. Experiments

In this section, we conduct experiments to evaluate the proposed method on three of the most challenging datasets: NTU-RGB+D [42], Kinetics-M [29], and SYSU-3D [36]. We first perform exhaustive ablation studies to examine the contributions of our proposed methods on the NTU-RGB+D dataset. Then the proposed model is tested and compared with state-of-the-art action recognition methods on three popular datasets.

4.1. Network architecture

We build a graph convolutional network as the baseline model with reference to Spatial Temporal Graph Convolutional Network (ST-GCN) [29], which is the first to apply graph convolutional networks to skeleton based action recognition. The structure of the baseline model is shown in Fig. 6. It is composed of nine ST-GC modules, and each module has a graph convolutional layer with kernel size three and a temporal convolutional layer with kernel size nine. It can learn both spatial and temporal information from the skeleton sequences. Before feeding input sequences into the baseline model, a batch normalization layer is used to normalize the data. The number of output channels in each layer is illustrated in Fig. 6. In more details, after 3th and 6th layers, we insert a Max Pooling on temporal domain respectively. Then a Global Average Pooling (GAP) layer is performed on the feature maps to obtain a 256 dimension feature vector. Finally, the feature vector of each sequence is fed into a SoftMax classifier to predict the action label. The proposed GCN with SGP and JCA is shown in Fig. 3. It has four graph convolutional layers which correspond to 4th, 7th, 8th and 9th layers of baseline model. There follows a temporal convolutional layer after each graph convolutional layer to extract temporal features. We employ a Max pooling with stride 2 on temporal dimension after each temporal convolutional layer. The SGP operations and JCA modules are plugged between graph convolutional layers. The other components, such as batch normalization layer, are the same as baseline model.

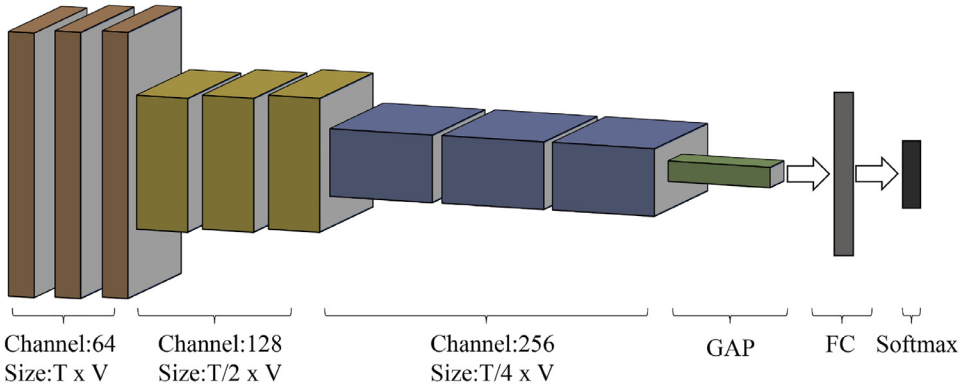


Fig. 6. The architecture of the baseline model with 9 graph convolutional layers. The feature dimensions are presented.

4.2. Datasets and training details

NTU-RGB+D Dataset [42]: NTU-RGB+D is a large dataset for human action recognition task with 60 action classes. It contains 56000 action clips captured from 40 volunteers in a constrained lab environment with 3 different horizontal angle cameras at the same height. Detected by Kinect V2 depth sensors, annotations of 3D joint coordinates of 25 major body joints are given for each frame. This dataset includes two evaluation metrics: Cross-Subject(CS) and Cross-View(CV). In the Cross-Subject evaluation, there are 40320 and 16560 clips for training and evaluation respectively. In this setting, 20 subjects from the dataset are used for training and the models are evaluated on clips from the remaining actors. In the Cross-View evaluation, there are 37920 and 18960 clips for training and evaluation respectively. Training samples in this setting are captured from camera view 2 and 3, the evaluation samples are all from camera view 1. We follow this convention settings and report the top 1 recognition accuracy on both evaluation metrics.

Kinetics-M [29]: Kinetics-M is an action video dataset presented by Yan et al. [29]. It contains 30 action classes seleceted from Kinetics. The joints coordinates are estimated from raw video clips with the public available realtime Openpose toolbox [47]. It contains 25000 clips for training and 1500 clips for evaluation. The training set is used to train our model, and we report the top 1 recognition accuracy on evaluation set.

SYSU-3D Dataset [36]: The SYSU-3D dataset is composed of 480 video clips and 12 different activities performed by 40 subject. The video clips from 20 subjects are used to train our model, and the video clips from the other subjects for evaluation.

Training Details: The proposed method is implemented with the PyTorch deep learning framework and the model is trained for 100 epochs on 2 TITANX GPUs. We employ Stochastic Gradient Descent (SGD) to train the model. The momentum and weight decay are set as 0.9 and 0.0001 respectively. The basic learning rate is set to 0.1 and is divided by 10 at 10, 50 and 90 of training epochs. We choose ReLU as the activation functions and set the dropout rate to 0.5 to avoid overfitting. According to the available GPU memory, the batch size of NTU-RGB+D, Kinetics-M, and SYSU-3D datasets are set to 16, 32, and 32, respectively. We train our model without any data augmentation.

4.3. Ablation study

To analyze the effectiveness and necessity of the proposed components, a series of ablation experiments on the NTU-RGB+D [42] dataset are conducted.

Table 1
Evaluating the effectiveness of proposed SGP scheme on the NTU-RGB+D dataset.

Method	Cross-Subject	Cross-View
Baseline(9L)	80.7%	88.9%
SGP^1(9L)	84.9%	91.6%
SGP^2(9L)	85.4%	92.7%
SGP^{3-1}(9L)	82.0%	91.1%
SGP^{3-2}(9L)	85.5%	93.1%
SGP^{3-3}(9L)	85.9%	93.3%
SGP^{3-3}(4L)	86.1%	93.1%

4.3.1. Evaluation of SGP scheme

To evaluate the effectiveness of the proposed SGP scheme, we design different groups of settings including: 1) SGP^1, 2) SGP^2 and 3) SGP^3. Besides, we compare three division approaches for the 3rd SGP operation to figure out how to extract high-level features more effectively. The details of SGP^i settings is described below.

- SGP^1: We insert the 1st SGP operation mentioned in Section 3.4 after 4th GC layer of the baseline model, and pool the original 25 joints to 10 nodes.
- SGP^2: The 1st and 2nd SGP operations are added after 3th and 6th GC layers of the baseline model, and pool the number of joints from 25 to 10 and 10 to 5, respectively.
- SGP^3: We insert the 1st, 2nd and 3rd SGP operations after 3th, 5th, 7th GC layers of the baseline model. Moreover, the 3rd SGP operation has three division approach as mentioned in Section 3.4. Denote the SGP^3 with the first, second, third division approach as SGP^{3-1}, SGP^{3-2} and SGP^{3-3} respectively.

The results of the above methods are listed in Table 1. From Table 1, we observe three points as follows. Firstly, it can be seen that all the models with SGP scheme outperform the baseline models both in Cross-Subject and Cross-View, which verifies that the SGP scheme is suitable for skeleton based action recognition by gradually expanding receptive fields to learn high-level features.

Secondly, the performance of models is generally higher with increasing number of SGP operations. Namely, SGP^3 is better than SGP^2, and SGP^2 is better than SGP^1. The proposed SGP can fuse the features from low-level nodes and obtain high-level features by means of weighted summation. What's more, convolving on the graph with fewer nodes can obtain global information. Thus, the information represented by nodes features are more global and high-level when the number of pooling operation increases. The improved accuracies demonstrate that these global and high-level information is significant for classification.

Finally, the SGP^{3-3} performs better than SGP^{3-1}, SGP^{3-2}. The result indicates that it's a better way to divide the motion features

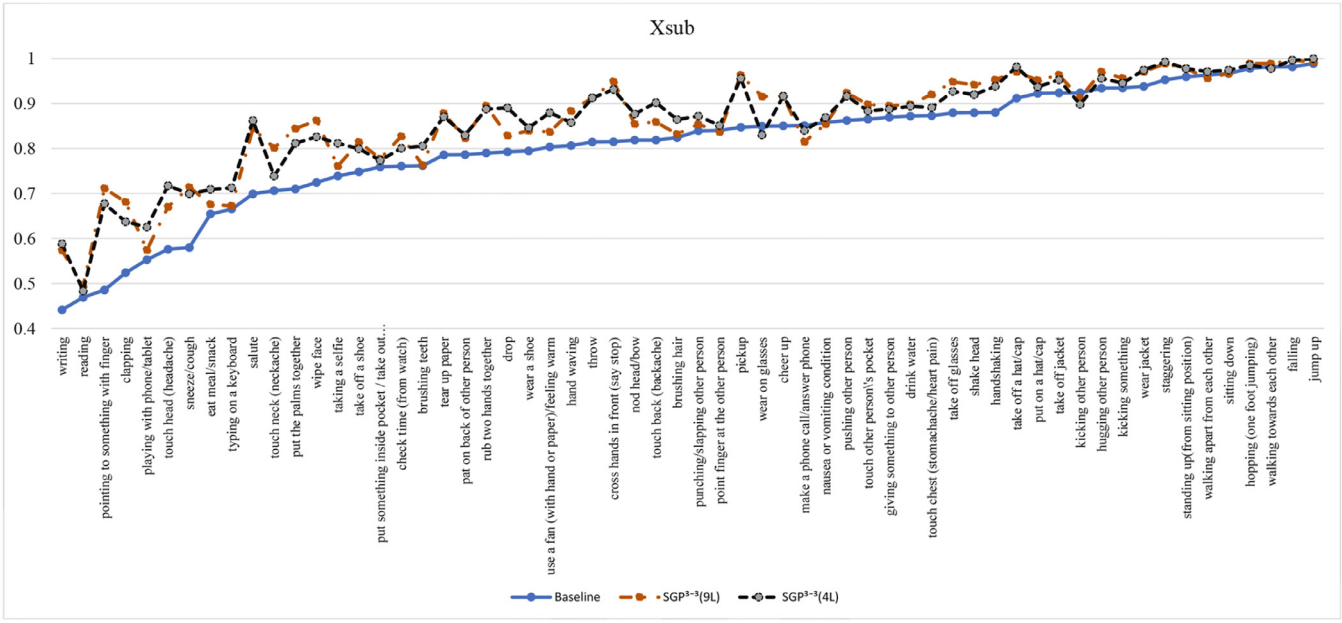


Fig. 7. Comparison results for each category on the NTU-RGB+D Cross-subject dataset. The horizontal axis represents the category and the vertical axis represents the accuracy. Better viewed in colour.

of five body parts to the motion features of upper body and lower body. Compared with SGP^{3-3}, the 3rd SGP operation of SGP^{3-1} and SGP^{3-2} are more specific, corresponding to left-upper body, right-upper body, legs and left-half body, right-half body, trunk respectively. Intuitively, the motion features of upper body and lower body are more discriminative for action classification since some action classes often only involve in the motions of hands and arms, while some action classes only involve in the motions of legs. However, the motion features of left-half body and right-half body are not discriminative for action classification since it should be classified as the same class if people perform the same action with different hands or legs. In the classified actions, there exist some actions which are improper to divide their motion into three parts. Thus, by intuitive analysis, it is a better way to divide joints into upper body and lower body. The experimental results coincide with the intuitive perception.

The proposed SGP scheme can accelerate the information transfer process on graphs. To prove the information transfer acceleration, we reduce the number of GC layers in SGP^{3-3} and compare the new model with the original SGP^{3-3}. We remove the 1st, 2nd, 3rd, 5th,6th GC layers of SGP^{3-3} and denote this four layers model as SGP^{3-3}(4L). To avoid confusion, denote the original SGP^{i} as SGP^{i}(9L). From Table 1, on the NTU-RGB+D dataset, the average accuracy of SGP^{3-3}(4L) and SGP^{3-3}(9L) are the same, both higher than the baseline model. Figs. 7 and 8 show the accuracy rates of 60 categories on the NTU-RGB+D dataset. As we can see, the performance of SGP^{3-3} (9L) almost coincides with that of SGP^{3-3} (4L), and accuracies on all categories are higher than the baseline model. The experimental results demonstrate that our method could decrease the number of graph convolution operations without affecting the accuracy,which means the SGP scheme can accelerate information transfer process.

We list some actions related to the interaction of the human body parts from the NTU-RGB+D dataset in Figs. 9 and 10. From the two figures, it can be seen that SGP^{3-3}(4L) achieves more than 9% improvements compared with the baseline model on these actions. Specifically, the classification accuracies of actions like “clapping”, “salute” and “touch head” are improved significantly. These actions are highly related with interactions of different parts of human body. The improvement shows that GCN equipped with SGP scheme can extract global motion information.

Table 2
Evaluating the performance of JCA module on the NTU-RGB+D dataset.

Method	Cross-Subject	Cross-View
Baseline	80.7%	88.9%
SGP^{3-3}(4L)	86.1%	93.1%
SGP^{3-3}+JCA^{1}(4L)	86.5%	93.3%
SGP^{3-3}+JCA^{2}(4L)	86.7%	93.6%
SGP^{3-3}+JCA^{3}(4L)	86.9%	93.8%

4.3.2. Evaluation of JCA module

To analyze the effectiveness of the joint-wise channel attention (JCA) module, we design three SGP+JCA models by inserting different numbers of JCA modules into SGP^{3-3}(4L). In Table 2, SGP^{3-3}+JCA^{1} (4L) plugs in a JCA module after the first graph convolutional layer of SGP^{3-3}(4L), SGP^{3-3}+JCA^{2} (4L) plugs in a JCA module after the first and second graph convolutional layers of SGP^{3-3}(4L) respectively, SGP^{3-3}+JCA^{3} (4L) plugs in a JCA module after the first, second and third graph convolutional layers of SGP^{3-3}(4L) respectively. Note that the SGP^{3-3}+JCA^{3} (4L) is the model we propose in this work. All of these models outperform the baseline model and SGP^{3-3} (4L). With the number of JCA modules increasing, the performance improves gradually both on Cross-Subject and Cross-View. The SGP^{3-3}+JCA^{3}(4L) achieves the best performance, which increases by 0.8% (CS) and 0.7% (CV) over SGP^{3-3} (4L). To analyze the computation efficiency of the proposed JCA module, we compare the training time of the baseline model and SGP^{3-3}(4L) with SGP^{3-3}+JCA^{3}(4L) under the same settings. As shown in Table 3, the training time of SGP^{3-3}+JCA^{3}(4L) and SGP^{3-3}(4L) are far less than that of the baseline model. The training time of SGP^{3-3}+JCA^{3}(4L) is 1.2 times of SGP^{3-3}(4L), which indicates that JCA module introduces little computation cost.

Additionally, to show the ability of our JCA module on classifying confusing human actions, we select some confusing human actions from the NTU-RGB+D dataset. Almost all these

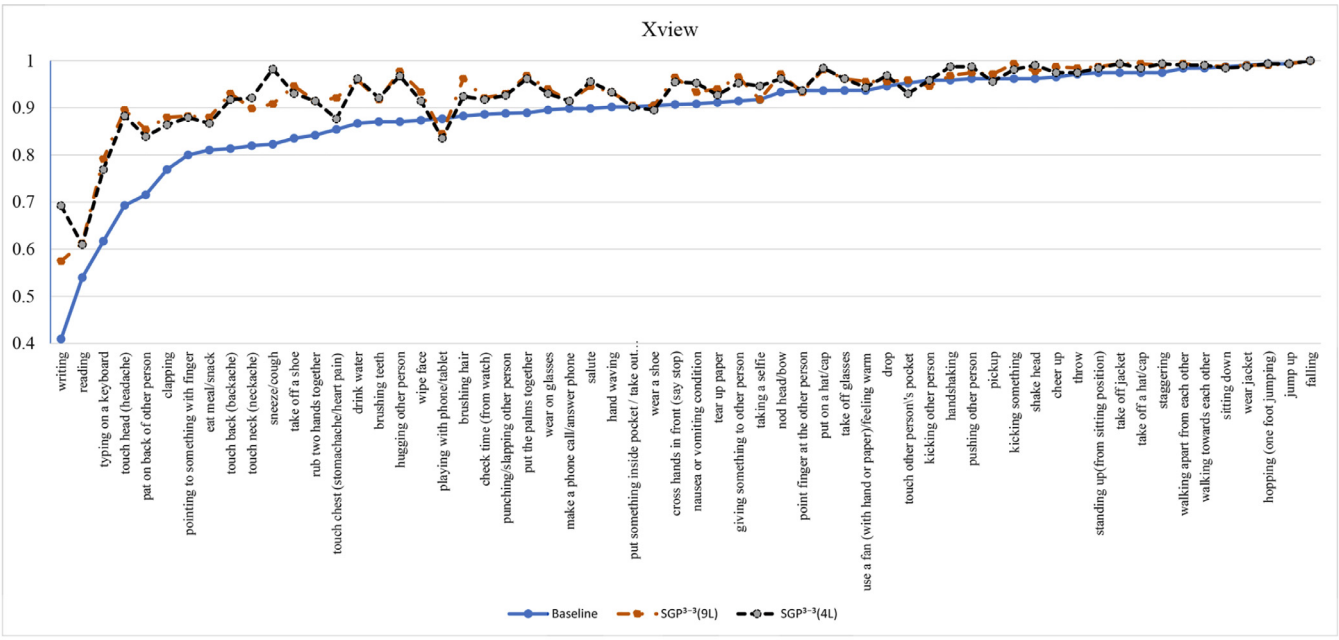


Fig. 8. Comparison results for each category on the NTU-RGB+D Cross-view dataset. The horizontal axis represents the category and the vertical axis represents the accuracy. Better viewed in colour.

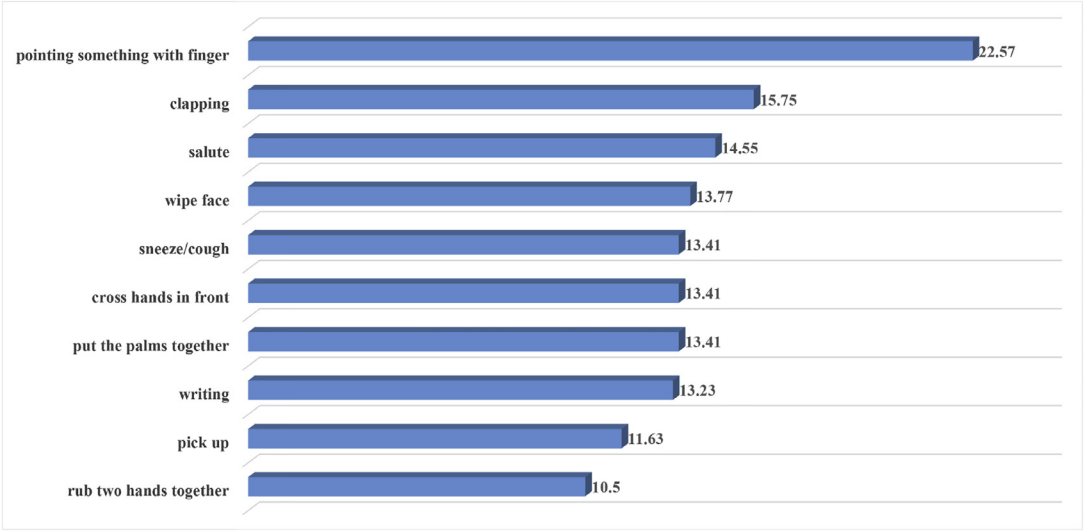


Fig. 9. The improved accuracies of $SGP^{3-3}(4L)$ compared with the baseline model on above categories in NTU-RGB+D Cross-Subject dataset.

Table 3
Training time of baseline model, $SGP^{3-3}(4L)$ and SGP^{3-3}+$JCA^3(4L)$ on the NTU-RGB+D dataset.

Method	Training time
Baseline	19 h 52 min
$SGP^{3-3}(4L)$	7 h 39 min
SGP^{3-3}+$JCA^3(4L)$	9 h 15 min

Table 4
Performance comparison for some confusing actions on the NTU-RGB+D Cross-Subject dataset.

Human actions	Baseline	$SGP^{3-3}(4L)$	SGP^{3-3}+$JCA^3(4L)$
clapping	52.4%	63.7%	72.9%
reading	46.9%	48.4%	56.0%
writing	44.1%	58.8%	65.1%
playing with phone/tablet	55.3%	62.6%	64.7%
point to something with finger	48.6%	67.8%	72.5%
sneeze/cough	58.0%	69.9%	64.9%
touch head	57.6%	71.7%	72.5%
put the palms together	71.0%	81.2%	92.8%
touch neck	70.7%	73.9%	80.8%
salute	69.9%	86.2%	91.3%

confusing actions are performed by hands such as "put palms together" "clapping", "reading" and "writing". From Table 4, compared with $SGP^{3-3}(4L)$, the model combined with JCA module can distinguish the confusing actions better. The accuracies of these actions are significantly improved except sneeze/cough.

We show the confusion matrix of "reading" and "writing" in Table 5. Compared with $SGP^{3-3}(4L)$, SGP^{3-3}+$JCA^3(4L)$ can discriminate these confusing actions better. This proves the proposed JCA module can focus on the key channels and enhance the discriminative information.

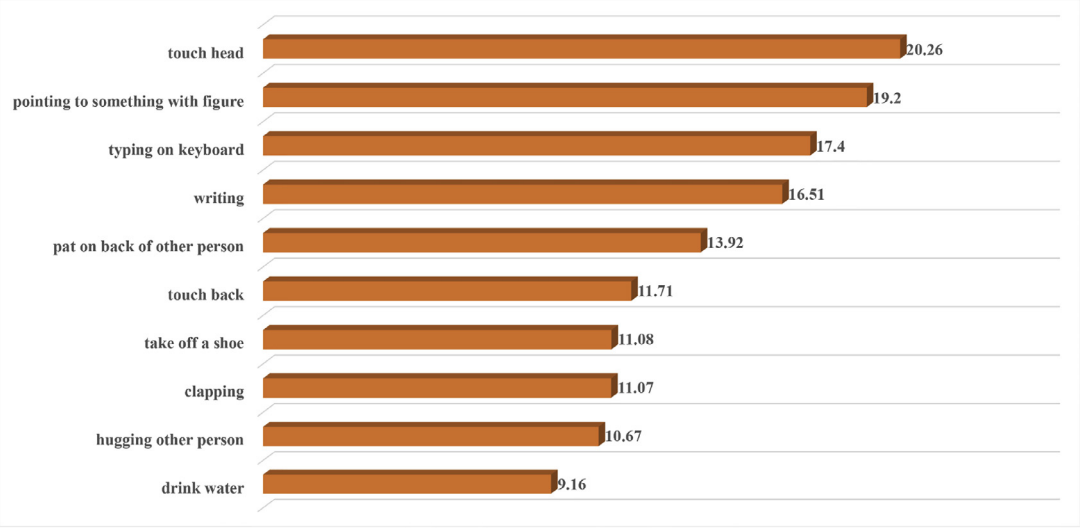


Fig. 10. The improved accuracies of SGP^{3-3}(4L) compared with the baseline model on above categories in NTU-RGB+D Cross-View dataset.

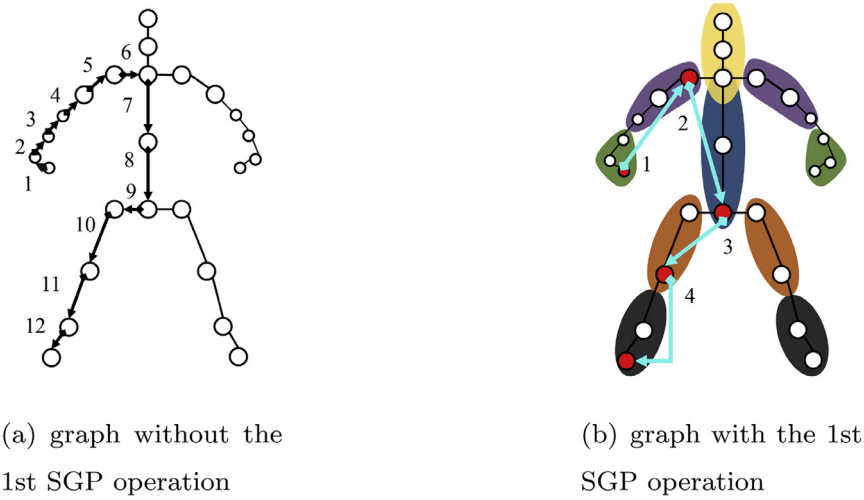


Fig. 11. Distances between "hand" and "right foot" with/without the 1st SGP operation. The numbers in picture (a) and (b) indicate lengths of paths from "hand" to "right foot". Pictures (a) and (b) are the graphs without and with the 1st SGP operation respectively. Better viewed in colour.

Table 5
The confusion matrix of "reading" and "writing" of SGP^{3-3}(4L)(represented by "S") and SGP^{3-3}+JCA^{3}(4L)(represented by "J") on NTU-RGB+D Cross-View dataset.

Human actions	reading(S)	writing(S)	reading(J)	writing(J)
reading	0.59	0.19	0.61	0.17
writing	0.13	0.68	0.06	0.69

Table 6
The recognition results, numbers of parameters and cost of multiplications with our methods on NTU-RGB+D dataset.

Method	Parameters	CS	CV
Baseline(9L)	2.98M	80.7%	88.9%
SGP^{3-3}(9L)	2.98M	85.9%	93.3%
SGP^{3-3}(4L)	2.49M	86.1%	93.1%

4.4. Computation cost evaluation

In this section, we first analyze the proposed SGP scheme's superiority in reducing model size from the view of information transfer. Then we analyze the computation cost reduction brought by the SGP scheme.

4.4.1. Information transfer acceleration

The SGP scheme accelerates information transfer process between distant joints. The graph convolution operation is often defined in first-order neighborhoods. This means the information of one node can only be passed to its adjacent nodes with one graph convolution operation. As shown in Fig. 11, in the original GCNs, the distance between "right hand" and "right foot" is 12. We need 12 graph convolution operations to transfer the information from "right hand" to "right foot". With the 1st SGP operation, the distance between these two joints becomes 4, which means we need only 4 graph convolution operations to transfer information between them. Meanwhile, the model size reduces because of the decreasing number of graph convolution operations.

To show the model size reduction brought by decreasing the number of graph convolution operations, we calculate the amount of parameters of baseline model, SGP^{3-3} (9L) and SGP^{3-3} (4L). From Table 6, the amount of parameters of the SGP^{3-3}(4L) is 2.49

Table 7
The recognition result, amount of parameters and cost of multiplications of ST-GCN, GCMVT and our method on the NTU-RGB+D dataset.

Method	Parameters	Multiplications	CS	CV
GCMVT [31]	5.75M	61.21B	84.2%	90.2%
ST-GCN [29]	3.03M	13.31B	81.5%	88.3%
SGP^{3-3}(4L)	2.49M	3.39B	86.1%	93.1%

Table 8
Comparison with current state-of-the-art methods on the NTU-RGB+D dataset. The accuracies are reported on both the Cross-subject (CS) and Cross-view (CV) benchmarks.

Method	CS	CV	Year
Dynamic Skeletons [36]	60.2%	65.2%	2015
HBRNN-L [40]	59.1%	64.0%	2015
Part-aware LSTM [42]	62.9%	70.3%	2016
ST-LSTM+Trust Gate [48]	69.2%	77.7%	2016
Two-Stream RNN [49]	71.3%	79.5%	2017
STA-LSTM [15]	73.4%	81.2%	2017
VA-LSTM [50]	79.2%	87.7%	2017
View invariant [18]	80.0%	87.2%	2017
ST-GCN [29]	81.5%	88.3%	2018
DPRL [30]	83.5%	89.8%	2018
GCMVT [31]	84.2%	90.2%	2019
Baseline	80.7%	88.9%	
SGP^{3-3}(4L)	86.1%	93.1%	
SGP^{3-3}+JCA^{3}(4L)	86.9%	93.8%	

Table 9
Comparison with current state-of-the-art methods on Kinetics-M dataset. We list top-1 classification accuracies.

Method	Accuracy	Year
ST-GCN [29]	72.4%	2018
Baseline	71.3%	
SGP^{3-3}(4L)	74.3%	
SGP^{3-3}+JCA^{3}(4L)	75.0%	

Table 10
Comparison with state-of-the-art methods on the SYSU-3D dataset.

Method	Accuracy	Year
Dynamic Skeletons [36]	75.5%	2015
LAFF(SKL) [51]	54.2%	2016
ST-LSTM(Tree) [48]	73.4%	2017
ST-LSTM(Tree)+Trust Gate [48]	76.5%	2017
DPRL [30]	76.9%	2018
Baseline	73.8%	
SGP^{3-3}(4L)	78.3%	
SGP^{3-3}+JCA^{3}(4L)	79.2%	

million, which is reduced by 16.4% compared with the baseline model and the SGP^{3-3} (9L). It can be seen that only four graph convolutional layers in our method are already sufficient to achieve better accuracy than the baseline model with nine layers.

4.4.2. Computation cost reduction

To further demonstrate the lightweight characteristic and computational efficiency of our method, we compare two GCN-based models ST-GCN [29] and GCMVT [31] with SGP^{3-3}(4L) on the NTU-RGB+D dataset and record their accuracies, model sizes and computation cost.

The experimental results are illustrated in Table 7. The following points can be seen from Table 7. (1) Our model achieves the best accuracies, which are about 5% higher than ST-GCN and 2% higher than GCMVT on average. (2) Our model with structure based graph pooling scheme is lightweight compared with other methods. The number of parameters is less than half of that of GCMVT, and also less than that of ST-GCN. (3) The computation cost of our model is smaller than other methods. Multiplication operations occupy the most computation cost during model training. Our method largely reduces the number of multiplication operations. On this aspect, GCMVT is 18 times and ST-GCN is almost 4 times of our model. These points indicate that our pooling scheme can not only improve the accuracy of graph convolutional network but also reduce the amount of model parameters and computation cost.

Theoretically, this computation reduction comes from two properties of SGP scheme. Firstly, SGP scheme expands the receptive fields of graph convolution kernels effectively. This allows our model to extract global motion features with fewer graph convolutional layers. By reducing the number of graph convolutional layers, the computation cost of the whole model is decreased. Secondly, SGP scheme reduces the graph size, i.e. the number of nodes in the graph. The computation cost involved in a graph is linear with the graph size. The computation cost involved in a graph containing M nodes is about M/N times that in a graph containing N nodes. The skeleton in the NTU-RGB+D dataset has 25 joints(nodes). When we apply the proposed SGP and reduce the number of nodes to 5, the computation cost is reduced by about 5 times.

4.5. Comparison with state-of-the-art

We demonstrate the effectiveness of our method on three large scale datasets and compare with state-of-the-art methods as well as our baseline model.

4.5.1. NTU-RGB+D

We compare our SGP^{3-3} (4L) and SGP^{3-3}+JCA^{3} (4L) models with current state-of-the-art methods on the NTU-RGB+D dataset with two recommended evaluation metrics: Cross-Subject and Cross-View. The comparison results with these state-of-the-art methods are list in Table 8. Our method SGP^{3-3}+JCA^{3} (4L) achieves the performance of 86.9% (CS) and 93.8% (CV). Compared with other current state-of-the-art GCNs based methods, our method outperforms GCMVT [31] by 2.7% (CS) and 3.6% (CV), which shows the effectiveness of the proposed model.

4.5.2. Kinetics-M

Table 9 presents the comparison performance with the stat-of-the-arts on the Kinetics-M dataset. Our SGP^{3-3}+JCA^{3} (4L) achieves 2.5% improvement over the ST-GCN [29] and achieves 3.6% improvement over our baseline model. The Kinetics-M is a relative new dataset, thus the methods that can be compared are limited.

4.5.3. SYSU-3D

On the SYSU-3D dataset, We compare our method with five state-of-the-art approaches. Table 10 shows that our SGP^{3-3}+JCA^{3} (4L) outperforms the latest method DPRL [30] over 2.3% and outperforms our baseline model over 5.4%.

5. Conclusion

In this paper, we propose a novel graph convolutional network with structure based graph pooling (SGP) scheme and joint-wise channel attention (JCA) modules. The SGP scheme gradually pools the human skeleton graph and expands the receptive fileds of the convolution kernel to learn high-level information of human body. Specially, the proposed SGP scheme can enhance the GCNs' ability to extract global motion features and reduce the computation cost at same time. We have proven that SGP can accelerate information transferring and thus reduce the amount of parameters

and computation cost. Joint-wise channel attention (JCA) module learns to selectively focus on discriminative joints of skeleton and pays different levels of attention to different channels. On three widely used datasets: NTU-RGB+D, Kinetics-M and SYSU-3D, our method outperforms all state-of-the-art methods. For the limitations of our method, the proposed SGP scheme is designed for skeleton based data and cannot be directly used for other data. Besides, the SGP scheme is a hand-crafted method, i.e. we have to design the approach to divide the original graph to subgraphs manually. Although we explore several approaches in our work, it may not be optimal for skeleton based action recognition. It's a better way to design an adaptive pooling scheme. In the future work, we will improve our model to make it more adaptive to the similar actions.

Acknowledgment

This work is partly supported by the National Key R&D Plan (Nos. 2017YFB1-002801 and 2016QY01W0106), the Natural Science Foundation of China (Nos. U1803119, U1736106, 61751212, 61721004, 61972397, and 61772225), the NSFC-General Technology Collaborative Fund for Basic Research (Grant No. U1636218), the Key Research Program of Frontier Sciences, CAS (Grant No. YZDJ-SSW-JSC040), Beijing Natural Science Foundation (Nos. JQ18018, L172051 and L182058) and the CAS External Cooperation Key Project. Bing Li is also supported by Youth Innovation Promotion Association, CAS.

References

[1] U. Gaur, Y. Zhu, B. Song, A. Roy-Chowdhury, A string of feature graphs model for recognition of complex activities in natural videos, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2595–2602.
[2] Z. Duric, W.D. Gray, R. Heishman, F. Li, A. Rosenfeld, M.J. Schoelles, C. Schunn, H. Wechsler, Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction, Proc. IEEE 90 (7) (2002) 1272–1289.
[3] M. Sudha, K. Sriraghav, S.G. Jacob, S. Manisha, et al., Approaches and applications of virtual reality and gesture recognition: a review, Int. J. Ambient Comput.Intell. (IJACI) 8 (4) (2017) 1–18.
[4] M. Firman, RGBD datasets: past, present and future, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 19–31.
[5] J. Zhang, W. Li, P.O. Ogunbona, P. Wang, C. Tang, RGB-D-based action recognition datasets: a survey, Pattern Recognit. 60 (2016) 86–105.
[6] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, C. Theobalt, Vnect: real-time 3d human pose estimation with a single RGB camera, ACM Trans. Graph. (TOG) 36 (4) (2017) 44.
[7] F. Han, B. Reily, W. Hoff, H. Zhang, Space-time representation of people based on 3d skeletal data: a review, Comput. Vis. Image Underst. 158 (2017) 85–105.
[8] L. Xia, C.-C. Chen, J.K. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2012, pp. 20–27.
[9] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1290–1297.
[10] M.A. Gowayyed, M. Torki, M.E. Hussein, M. El-Saban, Histogram of oriented displacements (hod): describing trajectories of human joints for action recognition, in: Twenty-Third International Joint Conference on Artificial Intelligence, 2013.
[11] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 588–595.
[12] C. Wang, Y. Wang, A.L. Yuille, Mining 3d key-pose-motifs for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2639–2647.
[13] J. Weng, C. Weng, J. Yuan, Spatio-temporal Naive-Bayes nearest-neighbor (ST-NBNN) for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4171–4180.
[14] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, et al., Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks., in: AAAI, vol. 2, 2016, p. 6.
[15] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[16] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal LSTM with trust gates for 3d human action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 816–833.
[17] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3d action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3288–3297.
[18] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recognit. 68 (2017) 346–362.
[19] E.P. Ijjina, K.M. Chalavadi, Human action recognition using genetic algorithms and convolutional neural networks, Pattern Recognit. 59 (2016) 199–212.
[20] M. Ma, N. Marturi, Y. Li, A. Leonardis, R. Stolkin, Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos, Pattern Recognit. 76 (2018) 506–521.
[21] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, S.J. Maybank, Asymmetric 3d convolutional neural networks for action recognition, Pattern Recognit. 85 (2019) 1–12.
[22] S. Qi, W. Wang, B. Jia, J. Shen, S.-C. Zhu, Learning human-object interactions by graph parsing neural networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 401–417.
[23] L. Fan, W. Wang, S. Huang, X. Tang, S.-C. Zhu, Understanding human gaze communication by spatio-temporal graph reasoning, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5724–5733.
[24] W. Wang, Y. Xu, J. Shen, S.-C. Zhu, Attentive fashion grammar network for fashion landmark detection and clothing category classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4271–4280.
[25] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, L. Shao, Learning compositional neural information fusion for human parsing, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5703–5713.
[26] H.-S. Fang, Y. Xu, W. Wang, X. Liu, S.-C. Zhu, Learning pose grammar to encode human body configuration for 3d pose estimation, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[27] W. Wang, X. Lu, J. Shen, D.J. Crandall, L. Shao, Zero-shot video object segmentation via attentive graph neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9236–9245.
[28] Z. Zheng, W. Wang, S. Qi, S.-C. Zhu, Reasoning visual dialogs with structural and partial observations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6669–6678.
[29] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, AAAI, 2018.
[30] Y. Tang, Y. Tian, J. Lu, P. Li, J. Zhou, Deep progressive reinforcement learning for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5323–5332.
[31] Y. Wen, L. Gao, H. Fu, F. Zhang, S. Xia, Graph CNNs with motif and variable temporal block for skeleton-based action recognition, in: Proc. the 33rd AAAI Conference on Artificial Intelligence, 2019.
[32] R. Ying, J. You, C. Morris, X. Ren, W.L. Hamilton, J. Leskovec, J.M. Antognini, J. Sohl-Dickstein, N. Roohi, R. Kaur, et al., Hierarchical graph representation learning with differentiable pooling, CoRR (2018).
[33] T.H. Nguyen, R. Grishman, Graph convolutional networks with argument-aware pooling for event detection, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[34] C. Si, W. Chen, W. Wang, L. Wang, T. Tan, An attention enhanced graph convolutional LSTM network for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1227–1236.
[35] O. Boiman, E. Shechtman, M. Irani, In defense of nearest-neighbor based image classification, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[36] J.-F. Hu, W.-S. Zheng, J. Lai, J. Zhang, Jointly learning heterogeneous features for RGB-D activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5344–5352.
[37] P. Koniusz, A. Cherian, F. Porikli, Tensor representations via kernel linearization for action recognition from 3d skeletons, in: European Conference on Computer Vision, Springer, 2016, pp. 37–53.
[38] P. Wang, C. Yuan, W. Hu, B. Li, Y. Zhang, Graph based skeleton motion representation and similarity measurement for action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 370–385.
[40] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.
[42] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu RGB+ D: a large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1010–1019.
[44] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, in: Advances in Neural Information Processing Systems, 2015, pp. 2224–2232.
[45] M. Niepert, M. Ahmed, K. Kutzkov, Learning convolutional neural networks for graphs, in: International Conference on Machine Learning, 2016, pp. 2014–2023.
[47] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.
[48] J. Liu, A. Shahroudy, D. Xu, A.C. Kot, G. Wang, Skeleton-based action recognition using spatio-temporal LSTM network with trust gates, IEEE Trans. Pattern Anal. Mach.Intell. 40 (12) (2018) 3007–3021.

[49] H. Wang, L. Wang, Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 499–508.
[50] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View adaptive recurrent neural networks for high performance human action recognition from skeleton data, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2117–2126.
[51] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, J. Lai, Real-time RGB-D activity prediction by soft regression, in: European Conference on Computer Vision, Springer, 2016, pp. 280–296.

Yuxin Chen recevied the B.S. degree from the Department of Automation, Beihang University, Beijing, China, in 2019. He is currently a postgraduate of the Institute of Automation, Chinese Academy of Sciences (CASIA). His main research area is skeleton-based action recognition.

Gaoqun Ma received the B.S. degree in computer science and technology, Northeastern University at Qinhuangdao in 2017. Currently, he is a MA.Sc student training in Beijing Jiaotong University. His research interests is action recognition.

Chunfeng Yuan received the PH.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2010. She was a visiting scholar at University of Adelaide, Australia in 2010, and in the Internet Media Group and the Media Computing Group at Microsoft Research Asia in 2016. She is currently a associate professor at the CASIA. Her research interests and publications range from statistics to computer vision, including sparse representation, deep learning, action recognition, and event detection.

Bing Li received the Ph.D. degree from the Department of Computer Science and Engineering, Beijing Jiaotong University, China, in 2009. He is currently an Associate Professor in the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include video understanding, color constancy, visual saliency, and web content mining.

Hui Zhang received the Masters degree from the Department of Electronic and Information Engineering, Beijing Jiaotong University, China, in 2007. She is currently pursuing a doctorate at the Institute of Information Engineering, Chinese Academy of Sciences. Her research interests include network behavior analysis, action recognition, and deep learning.

Fangshi Wang is a professor with Software Engineering School in Beijing Jiaotong University. She received the PhD degree from the School of Computer Science and Engineering, Beijing Jiaotong University, China, in 2007. Her research interests focus on Computer vision, Video analysis and semantic tag.

Weiming Hu received the Ph.D. degree from the Department of Computer Science and Engineering, Zhejiang University in 1998. From 1998 to 2000, he was a postdoctoral research fellow with the Institute of Computer Science and Technology, Peking University. Currently, he is a full professor in the Institute of Automation, Chinese Academy of Sciences. He has published more than 200 papers on international journals and conferences. His research interests include visual motion analysis and recognition of web objectionable information.