Skeleton-Based Action Recognition with Combined Part-Wise Topology Graph Convolutional Networks

Xiaowei Zhu, Qian Huang(✉), Chang Li, Jingwen Cui, and Yingying Chen

School of Computer and Information, Hohai University, Nanjing, China
huangqian@hhu.edu.cn

Abstract. Graph Convolutional Network (GCN) has achieved promising performance in skeleton-based action recognition by modeling skeleton sequences as spatio-temporal graphs. However, most existing methods only focus on the overall characteristics of the skeleton, thus lacking fine-grained exploration of human body parts semantics. In this paper, we propose a novel Combined Part-wise Topology Graph Convolutional Network (CPT-GCN), including SPT-GC, TPT-GC, and STPT-GC modules, to refine the spatio-temporal topology from the spatial, temporal, and spatio-temporal perspectives, respectively. Specifically, SPT-GC aggregates spatial features by combining global topology and partial correlations. TPT-GC combines the overall motion trend and the motion details of parts to capture temporal dynamics. STPT-GC establishes a spatio-temporal dependency, focusing on exploiting the implicit spatio-temporal information in motions. Ultimately, the effectiveness of CPT-GCN is demonstrated through experiments on two large-scale datasets: NTU RGB+D 60 and NTU RGB+D 120.

Keywords: Skeleton action recognition · Graph convolutional network · Part-wise topology · Spatio-temporal correlation

1 Introduction

As a major research topic of computer vision technology, human action recognition plays an important role in applications such as video surveillance, human-computer interaction and abnormal behavior detection [1,3,8,14,27]. Skeleton data is a compact and expressive modality that has less data volume compared with RGB or depth modality, and is insensitive to complex backgrounds and dynamic camera perspectives. Therefore, skeleton-based human action recognition technology has received widespread attention [12,13,25,31,32].

Early deep learning-based action recognition methods manually construct human skeleton coordinates into vector sequences or pseudo-images, and feed them into a recurrent neural network (RNN) or convolutional neural network (CNN) to predict action results [5,7,10,36]. Kim et al. [11] used a one-dimensional residual CNN to identify skeleton sequences based on directly-concatenated joint coordinates. Li et al. [18] constructed an adaptive tree-structured RNN, and Si et al. [28] proposed a novel attention-enhanced graph

Q. Liu et al. (Eds.): PRCV 2023, LNCS 14425, pp. 43–59, 2024.
https://doi.org/10.1007/978-981-99-8429-9_4

convolutional LSTM network called AGC-LSTM for human action recognition from skeleton data. However, these methods ignore the important property of human skeleton as a topological structure, and it is difficult to capture the spatio-temporal dependencies between joints.

Graph Convolutional Network (GCN) can efficiently handle non-Euclidean data such as graphs, and it can generalize convolutions from images to graphs of arbitrary size and shape. In recent years, more and more skeleton action recognition models use GCN-based methods to extract spatio-temporal features [4,7,16,22,25,26,34]. Yan et al. [34] manually defined the human body topology, and Shi et al. [25] learned the human body topology dynamically through adaptive graph convolution. They all focus on graph convolution on the global human body topology while ignoring body part information. For many actions, such as clapping and throwing, the motion characteristics of parts are more important. Thakkar et al. [30] is the first to split the human skeleton into different parts for graph convolution. Wang et al. [33] proposed adaptive multi-part graph convolution to learn the spatial correlation between parts based on the self-attention mechanism. However, the topology of the human skeleton has not been fully utilized, and we construct the more refined local topology to extract more detailed features.

In this paper, we will further model the human skeleton topology from the three dimensions of spatial, temporal, and spatio-temporal based on human body parts. We then propose a novel network named Combined Part-wise Topology Graph Convolutional Networks (CPT-GCN), which focuses on exploring fine-grained features and capturing intrinsic spatio-temporal correlations. Specifically, we propose three modules, SPT-GC, TPT-GC and STPT-GC, to perform graph convolution based on locally refined topology. SPT-GC establishes specific global and local topologies in different channels, taking into account both global and local information to capture the spatial connections of joints in more detail. TPT-GC reasonably changes the receptive field of temporal convolution to extract the motion trend and motion details of the whole and part of the action. STPT-GC focuses on extracting the implicit spatio-temporal association information in the skeleton sequence, and establishes the part-enhanced spatio-temporal association topology. Combining the above three modules, our network dynamically aggregates high-dimensional features and achieves excellent performance on large-scale datasets.

Combining these efforts above, our main contributions are summarized as follows:

- Our proposed SPT-GC refines the spatial topology based on body parts by fusing global and local topology, which extracts more fine-grained spatial features.
- We propose the spatio-temporal module, including TPT-GC and STPT-GC, which establishes a specific temporal correlation topology and spatio-temporal correlation topology, and effectively extracts the temporal and spatio-temporal correlation of parts and joints.

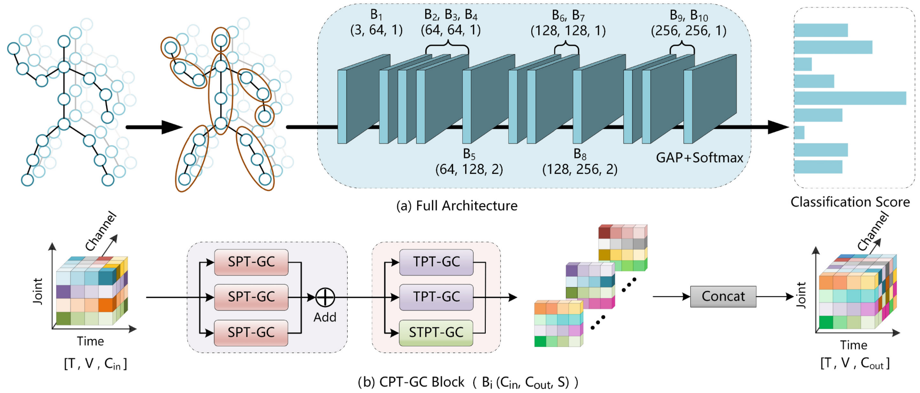
- We propose a novel action recognition model CPT-GCN based on skeleton data. It accurately captures the relationship between and within parts, and effectively aggregates the spatial, temporal and spatio-temporal information of skeleton data.
- We conduct experiments on two widely-used datasets: NTU RGB+D [24] and NTU RGB+D 120 [19], on which our proposed method outperforms state-of-the-art approaches.

2 Related Work

2.1 Skeleton-Based Action Recognition

With the development of deep learning technology, deep learning methods have gradually replaced traditional manual feature methods. The mainstream methods can be divided into three categories according to the network architecture: convolutional neural network (CNN), recurrent neural network (RNN) and graph convolutional network. (GCN).

CNN-based method usually converts the skeleton data into a pseudo-image according to the manually designed conversion rules. RNN-based methods usually extract frame-level skeleton features, represent skeleton data as sequential data with predefined traversal rules [4,18]. However, human skeleton is a natural graphical structure, and GCN has obvious advantages in processing graph-structured data. Yan et al. was the first to use GCN to model human skeleton, proposing Spatio-temporal Graph Convolutional Network (ST-GCN). They build joint connection edges based on the natural connections of the human body,



(a) Full Architecture

(b) CPT-GC Block (B_i (C_{in}, C_{out}, S))

Fig. 1. The overview of the proposed CPT-GCN model. The entire combined part-wise topology graph convolutional block is represented as $B_i(C_{in}, C_{out}, S)$. C_{in}, C_{out} and S denote the number of input channels, the number of output channels and the stride, respectively. There are a total of 10 blocks. GAP represents the global average pooling.

and add temporal associations for the same joints in consecutive frames, constructing a skeletal spatio-temporal graph [34]. Shi et al. proposed an adaptive graph convolution network (AGCN), which uses the self-attention mechanism to change the topology of human skeletons and adaptively learns the connection between the original disconnected skeletons [25,26]. Liu et al. introduced a multi-scale graph topology to achieve multi-scale joint relationship modeling [21]. Cheng et al. proposed Shift-GCN [7], replacing the traditional convolution operator with the shift convolution operator, using shifted graph convolution. The CTR-GCN proposed by Chen et al. [5] designs channel-wise topology graphs to explore more possibilities for feature learning in different channels.

2.2 Partial Graph Convolution in Skeleton-Based Action Recognition

A complete action can be regarded as composed of different postures of human body parts. For example: in the process of clapping, the clapping of the palm plays a key role in the whole action, while the waving of the arm plays an auxiliary role. Previous studies [7,21,25,26,34] mostly learn the global features of actions based on the whole skeleton, ignoring the important contribution of local features to actions. Thakkar et al. [30] is the first to split the human skeleton into different parts for graph convolution, which effectively improves performance of recognition. Wang et al. [33] proposed adaptive multi-part graph convolution to learn the spatial correlation between parts based on the self-attention mechanism. Zhu et al. [38] focused on fusing global and local features from a spatial perspective, effectively aggregating multi-level joint features by constructing a topology based on bodyparts.

3 Methods

In this section, we first introduce the construction of skeletal spatio-temporal graph and conventional graph convolution. Then we elaborate the modeling strategies of part-wise spatial topology and spatio-temporal topology respectively. Finally, as shown in Fig. 1, we present the full model structure of the proposed Combined Part-wise Topology Graph Convolutional Networks model named CPT-GCN.

3.1 Preliminaries

Graph Construction. A full action consists of multiple frames containing different samples. We construct spatio-temporal skeleton graphs to describe the structured information between nodes along the spatial and temporal dimensions. The complete spatio-temporal skeleton graph is established based on the natural connections of the human body structure and the connection of consecutive frames, so it contains the connection edges between joints and the connection edges between frames. The graph is defined as $\mathcal{G} = (\mathcal{X}, \mathcal{V}, \mathcal{E})$. $\mathcal{X}$ denotes the feature set of vertices, which is represented as a matrix $X \in R^{C \times V \times T}$, there are V vertices, T frames and C channels. $\mathcal{V} = \{v_1, v_2, ..., v_V\}$ denotes the vertex set. $\mathcal{E}$ is the set of edges, reflecting the connection strength between vertices.

Graph Convolution. After the skeleton spatio-temporal feature map is constructed, we weight and sum the skeleton points in the input feature map with the features of their corresponding neighbor points to obtain the output feature map. The graph convolution implementation of feature maps can be intuitively formulated as:

$$f_{out} = \sum_{s}^{S} W_s \cdot f_{in} \cdot A_s \tag{1}$$

where f_{in} and f_{out} denote the input and output feature maps. S denotes the sampling area of the spatial dimension. A_s and W_s denote the adjacency matrix and weight function under the sampling area s.

3.2 Part-Wise Spatial Modeling

Almost any action is composed of sub-actions of different parts, and the difference mainly lies in the correlation between parts and the contribution of parts to the whole action. For example, clapping can be decomposed into the action of two palms and arms, and nodding can be regarded as the action of the head. Thus, optimizing the topology of skeleton based on human body parts can more accurately obtain the dependencies between joints.

Most of the previous studies explored the global features of the skeleton, and learned the spatial relationship of the skeleton through the natural connection of the human body or the attention mechanism [25,26,32,34], which will generate a lot of redundant information, and the spatial topology shared by each channel is also not optimal. Existing part-based models usually aim to extract features from body parts individually or only focus on discovering the importance of different body parts [29,35]. However, we take full account of inter-part dependencies and intra-part differences, and construct a refined part-wise topology for each channel.

Before performing GCN, body part correlations need to be modeled. Specifically, we divide the human body into 8 parts, which are head, body, two arms, two palms and two legs. The input features $X \in R^{C \times V \times T}$ is aggregated according to the proposed part division strategy, which is formulated as:

$$X_i^{part} = Concat(\{X_j \mid j \in L(i)\}) \quad i = 1, 2, ..., P \tag{2}$$

where P denotes the number of parts, $Concat(\cdot)$ denotes the splicing function, $L(i)$ denotes the set of joint numbers corresponding to the i$_{th}$ part, and X_i^{part} denotes the feature of the i$_{th}$ part after aggregation.

Parts Correlation Modeling. In order to obtain the best dependencies between parts, we propose the modeling strategie $\mathcal{M}(\cdot)$ to model the part dependencies.

Since each joint contributes to the corresponding body part, we perform an average pooling operation on the joints inside the part. In addition, in order to reduce the computation cost, we utilize linear transformations $\psi(\cdot)$ and $\varphi(\cdot)$ to reduce the feature dimension before the local topology modeling. $\mathcal{M}(\cdot)$ needs to calculate the distance of the channel dimension between different parts, and utilizes the nonlinear transformation of the distance to represent the correlation between parts, which is formulated as:

$$\mathcal{M}(i,j) = \sigma(\psi(AvgPool_{ST}(X_i^{part})) - \varphi(AvgPool_{ST}(X_j^{part}))) \tag{3}$$

where $AvgPool_{ST}(\cdot)$ denotes the average pooling in both spatial and temporal dimensions. $\mathcal{M}(i,j)$ is the modeling strategy, and its value denotes the correlation between parts i and j.

Part-wise Topology Modeling. The part correlation graph obtained by $\mathcal{M}(\cdot)$ represents the correlation between parts and cannot be directly applied to the human skeleton graph, so it needs to be mapped to joints relation graph through a mapping function. According to the relationship between the various parts obtained, the part correlation features are first connected into a whole vertex matrix, which is formulated as:

$$G_{part} = Concat(\{Concat(\{\mathcal{M}(i,j) \mid j = 1,2,...,P\}) \mid i = 1,2,...,P\}) \tag{4}$$

where G_{part} denotes the spliced inter-part relationship graph. It expresses different part correlations on each channel. But in fact, the joints within a part do not share weights, so the topology needs to be refined while mapping. We optimize the topology through learnable bias and linear transformation, which is formulated as:

$$G_s^{local} = \phi(\mathcal{R}(G_{part}) + B_0) \quad s = 1,2,...,S \tag{5}$$

where $\mathcal{R}(\cdot)$ denotes the mapping function, it maps the part association graph to the joint association graph. $\phi(\cdot)$ denotes the linear transformation function. B_0 denotes the positional bias of the channel and joint, which is a learnable parameter. G_s^{local} is the feature map based on body parts.

Spatial Part-wise Topology Graph Convolution(SPT-GC). Local topology captures both part correlations and intra-part differences. On this basis, a global topology is introduced to perform adaptive learning driven by data to capture the global spatial characteristics of actions. Our proposed CPT-GC is more flexible, which combines global and local topology to more accurately obtain the correlation between human skeletons. A gating mechanism α is introduced in the process of fusing the global graph and the individual refined graph to control the difference in the contribution of required parts and joints in different sampling regions. Finally, the graph convolution can be completed by performing Einstein summation of the part-wise topology and the input features in the spatial dimension.

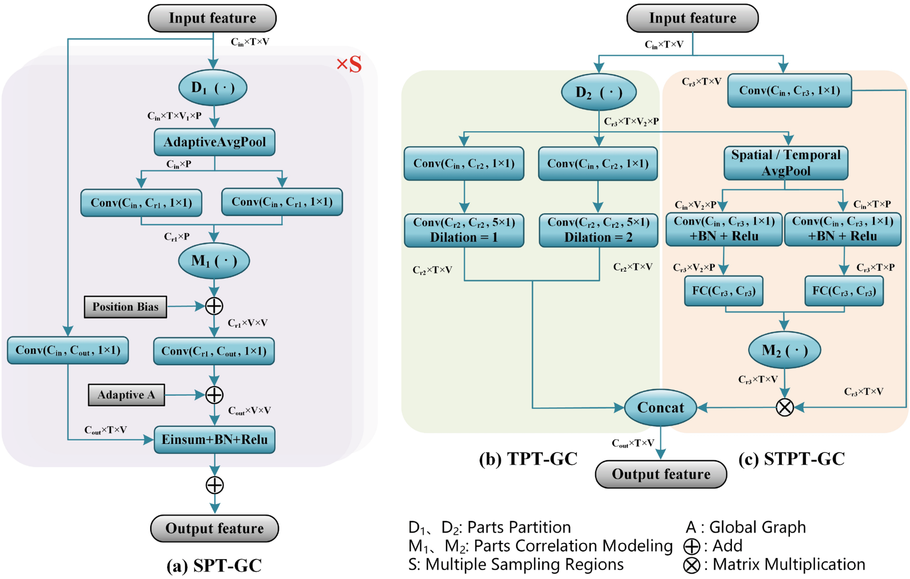

Input feature
$C_{in}\times T\times V$
×S
D_1 (·)
$C_{in}\times T\times V_1\times P$
AdaptiveAvgPool
$C_{in}\times P$
Conv(C_{in}, C_{r1}, 1×1)
Conv(C_{in}, C_{r1}, 1×1)
$C_{r1}\times P$
M_1 (·)
Position Bias
$C_{r1}\times V\times V$
Conv(C_{in}, C_{out}, 1×1)
Conv(C_{r1}, C_{out}, 1×1)
Adaptive A
$C_{out}\times V\times V$
$C_{out}\times T\times V$
Einsum+BN+Relu
Output feature
(a) SPT-GC
Input feature
$C_{in}\times T\times V$
D_2 (·)
$C_{r3}\times T\times V$
Conv(C_{in}, C_{r3}, 1×1)
$C_{r3}\times T\times V_2\times P$
Conv(C_{in}, C_{r2}, 1×1)
Conv(C_{in}, C_{r2}, 1×1)
Spatial / Temporal AvgPool
$C_{in}\times V_2\times P$
$C_{in}\times T\times P$
Conv(C_{r2}, C_{r2}, 5×1) Dilation = 1
Conv(C_{r2}, C_{r2}, 5×1) Dilation = 2
Conv(C_{in}, C_{r3}, 1×1) +BN + Relu
Conv(C_{in}, C_{r3}, 1×1) +BN + Relu
$C_{r2}\times T\times V$
$C_{r2}\times T\times V$
$C_{r3}\times V_2\times P$
$C_{r3}\times T\times P$
FC(C_{r3}, C_{r3})
FC(C_{r3}, C_{r3})
M_2 (·)
$C_{r3}\times T\times V$
$C_{r3}\times T\times V$
Concat
$C_{out}\times T\times V$
(b) TPT-GC
Output feature
(c) STPT-GC
D_1、D_2: Parts Partition
M_1、M_2: Parts Correlation Modeling
S: Multiple Sampling Regions
A : Global Graph
⊕ : Add
⊗ : Matrix Multiplication

Fig. 2. Model architecture of CPT-GCN block. It consists of three modules: SPT-GC, TPT-GC, and STPT-GC. $D_1(\cdot)$, $D_2(\cdot)$ denote parts partition function. $M_1(\cdot)$, $M_2(\cdot)$ denote parts correlation modeling function. *FC* denotes the fully connected layer. *BN* denotes Batch Normalization. *Relu* is the activation function.

GCN will dynamically update the global and local topology during the inference process to capture the features of the previously disconnected joints. Therefore, Eq. 1 is modified into the following form:

$$f_{out} = \sum_{s}^{S} W_s \cdot f_{in} \cdot (G_s^{global} + \alpha G_s^{local}) \tag{6}$$

where G_s^{global} is the global topology, which is initialized with the natural connection of the human skeleton, and changed by adaptively learning the correlation of actions.

The complete SPT-GC module is shown in Fig. 2 (a). We first divide the bone input feature X_{in} into parts, and then perform adaptive average pooling on the aggregated features. After that, they are respectively input to two convolutional layers with a convolution kernel of 1×1 for dimensionality reduction. After part-wise modeling, the associated topology graph of the part is obtained. Then it needs to be mapped to joint topology and fused with the global topology. In addition, multiple sampling regions S are set to learn semantic information at different levels.

3.3 Part-Wise Spatio-Temporal Modeling

The skeleton feature map composed of human action sequences contains rich spatio-temporal semantics, and there is actually a certain relationship between spatial and temporal information. We propose novel TPT-GC and STPT-GC for extracting temporal and spatio-temporal semantic information of action.

Based on the temporal perspective of the action, a complete action is composed of multiple sub-actions, such as squatting, bouncing, jumping forward, and standing can constitute a complete long jump action. TCN [2] learns the associated information between sub-actions or the trajectory of a complete action by setting convolution kernels of different sizes. But in fact, the sub-actions composed of different actions have different periods. Some actions pay more attention to long-term motion trends, others actions need to rely on short-term motion details to distinguish. Our designed TPT-GC contains different convolutional dilation coefficients, which focus on capturing long-term motion trends and short-term motion details, respectively.

Most of the previous methods extract the features of space and time separately, ignoring the internal relationship of time and space in the action. In fact, if we can extract the correlation between non-corresponding joints between frames, it will surely improve the accuracy of action recognition. Our proposed STPT-GC is used to capture spatio-temporal correlation features, and the effectiveness is verified in ablation experiments, as shown in Table 2.

In addition, the sub-actions that occur in different human body parts are also different. The arms and thighs may dominate the motion trend of this action, or the hands control the motion details of a certain action. It is obvious that adding part information helps to promote the learning of motion paterns. Therefore, we also introduced the concept of parts in the spatio-temporal modeling, and constructed the refined temporal and spatio-temporal topology respectively, achieving the part-enhanced effect.

Temporal Part-wise Topology Graph Convolution(TPT-GC). Inspired by Multi-scale Temporal Convolution [21], we design a part-based temporal modeling module for finer-grained extraction of joint motion trends and motion details. The part division strategy of Eq. 2 is used to aggregate the joint features of body parts. In order to reduce the computational complexity of the model, we utilize the $\psi(\cdot)$ linear transformation function to reduce the feature dimension. We set two convolution branches with different expansion coefficients in parallel to expand the neighborhood learned by graph convolution and extract semantic information at different levels of actions. The TPT-GC module is shown in Fig. 2 (b), which is formulated as:

$$f_{out}^{1}(i) = \sum_{k}^{K} W_1 \cdot \psi(f_{in}(i+k)) \quad i = 1, 2, ..., T \tag{7}$$

$$f_{out}^{2}(i) = \sum_{k}^{K} W_2 \cdot \psi(f_{in}(i+2k)) \quad i = 1, 2, ..., T \tag{8}$$

where f_{out}^1, f_{out}^2 denote the output feature obtained by the two branches. W_1, W_2 denote the weight corresponding to the convolution. K is the size of the convolution kernel in the time dimension.

Spatio-temporal Part-wise Topology Graph Convolution(STPT-GC). In order to obtain the inherent spatio-temporal correlation information of the action, we designed a novel spatio-temporal modeling module, which is also guided by the part information to establish the more refined spatio-temporal correlation topology. The STPT-GC module is shown in Fig. 2 (c). Specifically, STPT-GC relies on the spatio-temporal correlation topology to obtain spatio-temporal correlation information, and needs to construct a spatial correlation graph and a temporal correlation graph first. It uses the same part division strategy to aggregate joints features of the parts, and then aggregates the temporal and spatial information respectively through the average pooling operation. A linear transformation function is then used to reduce the temporal and spatial feature dimensions. It is formulated as:

$$G_{out}^S = W_S \cdot \sigma(\phi_1(AvgPool_S(f_{in})))) \tag{9}$$

$$G_{out}^T = W_T \cdot \sigma(\phi_2(AvgPool_T(f_{in}))) \tag{10}$$

where G_{out}^S and G_{out}^T denote the spatial and temporal correlation graphs, respectively. $AvgPool_S(\cdot)$ and $AvgPool_T(\cdot)$ denote the average pooling operation on the spatial and temporal dimensions, respectively. $\phi_1(\cdot)$ and $\phi_2(\cdot)$ denote the linear transformation function. $\sigma(\cdot)$ denotes the activation function. We add learnable parameters W_S and W_T to assist in learning the spatio-temporal features of actions, and then use the Kronecker product to model the spatio-temporal correlation topology. It is formulated as:

$$G_{out}^{ST} = \sigma(G_{out}^S \times G_{out}^T) \tag{11}$$

where G_{out}^S and G_{out}^T denote the spatial and temporal correlation graphs, respectively. G_{out}^{ST} denotes the obtained spatiotemporal correlation topology. Our proposed STPT-GC is parallel to TPT-GC. The output features of TPT-GC and STPT-GC are concatenated after spatio-temporal topological graph convolution. It is formulated as:

$$f_{out}^3 = W \cdot \phi_3(f_{in}) \cdot G_{out}^{ST} \tag{12}$$

$$f_{out} = Concat(f_{out}^{(i)}) \quad i = 1, 2, ..., N^{branch} \tag{13}$$

where f_{out}^3 denotes the output feature of the STPT-GC module. $\phi_3(\cdot)$ denotes the linear transformation function. $f_{out}^{(i)}$ denotes the output feature of the i_{th} branch. f_{out} denotes the output features after N^{branch} branches are cascaded. It can be understood that the first part of the channel represents the temporal characteristics of the action, and the latter part of the channel represents the spatiotemporal correlation characteristics of the action. The joints of each part can be restored to the original feature dimension through mapping and splicing strategies.

3.4 Model Architecture

We synthesize three modules of SPT-GC, TPT-GC and STPT-GC to construct a powerful graph convolutional network CPT-GCN for skeleton-based action recognition. The overall architecture is shown in Fig. 1 (a), which mainly consists of 10 basic blocks and a classification layer. The output channels of each block in the middle are 64, 64, 64, 64, 128, 128, 128, 256, 256, and 256. The residual network is connected between blocks [9], and finally perform global average pooling and softmax classification to obtain behavior prediction results.

Specifically, each individual block contains a spatial model and a spatio-temporal model, which are responsible for extracting spatial features and spatio-temporal joint features in skeleton information, respectively. As shown in Fig. 1 (b).

Spatial Modeling. The spatial model is mainly composed of SPT-GC modules, and three SPT-GCs are used in parallel to extract semantic information at different levels between parts and joints, as shown in Fig. 2 (a). For a single SPT-GC, first utilizes channel reduction rate r1 to compact representations, uses temporal and intra-part spatial pooling to aggregate features. After that, SPT-GC conducts pair-wise subtraction and activation, then fused with the global map. Finally, the graph convolution is completed to obtain the output feature map, as shown in Eq. 6.

Spatio-temporal Modeling. We demonstrate through ablation experiments that the spatio-temporal model with three branches has better performance. Among them, TPT-GC occupies two branches and STPT-GC occupies one branch, as shown in Fig. 2 (b) and (c).

TPT-GC first uses the channel reduction rate r2 to compress the channel information, and constructs two temporal convolutional layers of different scales to increase the receptive field, which are used to extract the motion trend and motion details of the action respectively.

STPT-GC aggregates temporal and spatial information through average pooling operations, and uses the channel reduction rate of r3 to reduce computational complexity. We use the Kronecker product to model spatio-temporal association topology. Finally, it performs a dot product of the compressed input features with the spatio-temporal correlation topology to complete the graph convolution, which can extract the spatio-temporal correlation information of the action.

4 Experiments

4.1 Datasets

NTU RGB+D. NTU RGB+D (NTU-60) [24] is currently the most widely used large-scale action recognition dataset, containing 60 action categories and 56,000 action clips. The clips were captured by three KinectV2 cameras with different perspectives and performed by 40 volunteers. Each sample contains one action and is guaranteed to have at most 2 subjects. The skeleton information consists of the 3D coordinates of 25 body joints and the corresponding action category labels. NTU-60 recommends two benchmarks [24]: Cross-View Evaluation (X-View) split according to different camera views and Cross-Subject Evaluation (X-Sub) split according to different subjects.

NTU RGB+D 120. NTU RGB+D 120 (NTU-120) [19] extends NTU-60 with a larger scale. It contains 120 action categories and 114,480 action clips. The clips were performed by 106 volunteers in 32 camera setups. NTU-120 also recommends two benchmarks [19]: the first is Cross-Subject Evaluation(X-Sub), which is the same cross-subject evaluation as NTU-60. The other is Cross-Setup Evaluation (X-Set), which splits training and test samples based on the parity of camera setup IDs.

4.2 Training Details

All experiments are conducted on one RTX 3070 TI GPU with the PyTorch deep learning framework. We use the stochastic gradient descent(SGD) with Nesterov momentum(0.9) as the optimizer and the cross-entropy as the loss function. Weight decay is 0.0004. The initial learning rate is set to 0.1 and a warmup strategy [9] is used in the first 5 epochs to make the training procedure more stable. The batch size is 32. The learning rate is divided by 10 at the 35_{th} epoch and 55_{th} epoch. The training process is ended at the 70_{th} epoch.Since the number of frames of the samples is not consistent, we uniformly downsample the frames to 64 frames. In addition, we adopt the data preprocessing strategy of [21] for the input skeleton features.

4.3 Ablation Studies

In this section, we use the X-Sub benchmark of the NTU-60 to verify the effectiveness of proposed modules in CPT-GCN.

Effectiveness of TPT-GC and STPT-GC. In order to test the performance of the space-time model proposed in Sect. 3.3 and obtain its optimal branch configuration, we conduct experiments on TPT-GC and STPT-GC with different branch numbers. We adopt ST-GCN [34] as the baseline method and replace the temporal module of the baseline model with the proposed spatio-temporal model. The specific ablation experiment configuration and results are shown in Table 1. The experimental results in the table show that the spatio-temporal model with two TPT-GCs and one STPT-GC branch configuration has better performance.

Model Configuration Exploration. As mentioned in Sect. 3.4, our proposed CPT-GCN contains three different modules, namely SPT-GC, TPT-GC and STPT-GC. We manually remove or only keep any kind of modules to test the parameter cost and model performance of different configurations of CPT-GCN. Additionally, we adopt ST-GCN [34] as the baseline method, which does not use any of these three modules.

Table 1. Comparison of the validation accuracy of spatio-temporal model with different settings.

Methods	Configuration	Acc(%)
Baseline	–	84.3
CPT-GCN (w/o SPT-GC)	TPT-GC + STPT-GC	87.5
	TPT-GC + 2STPT-GC	87.6
	TPT-GC + 3STPT-GC	86.6
	2TPT-GC + STPT-GC	**88.2**
	2TPT-GC + 2STPT-GC	88.0
	2TPT-GC + 3STPT-GC	87.1
	3TPT-GC + STPT-GC	87.7
	3TPT-GC + 2STPT-GC	87.4
	3TPT-GC + 3STPT-GC	86.9

Table 2. Comparison of the validation accuracy of CPT-GC with different settings.

Methods	SPT-GC	TPT-GC	STPT-GC	Param	Acc(%)
Baseline	–	–	–	1.27M	84.3
CPT-GCN	✓	✗	✗	2.30M	88.8
	✗	✓	✗	1.52M	87.5
	✗	✗	✓	1.45M	87.2
	✓	✓	✗	2.47M	89.1
	✓	✗	✓	2.40M	88.9
	✗	✓	✓	1.62M	88.2
	✓	✓	✓	2.57M	**89.5**

The specific ablation experiment configuration and results are shown in Table 2. The experimental results in the table show that although our proposed SPT-GC module introduces some additional parameters, it can effectively improve the performance of the model. The TPT-GC and STPT-GC modules have a significant effect on improving the performance of the model under the premise that a small number of parameters are required. The combination of the three modules of SPT-GC, TPT-GC and STPT-GC is the optimal configuration of this model. Under this configuration, CPT-GCN bring improvements of +5.2% over the baseline method on the X-Sub benchmark.

4.4 Comparison with the State-of-the-Art

Most state-of-the-art methods employ a multi-stream fusion framework to enrich semantic information. Our proposed method adopts the same strategy as [5,7,26] to generate four data modalities, namely joint, bone, joint motion and bone motion, and fuse the prediction scores of the four modalities.

We compare the final model with state-of-the-art skeleton-based action recognition methods on the NTU-60 and NTU-120 datasets. The results are shown in Tables 3 and 4. These methods for comparison include RNN-based methods [17,20,24], CNN-based methods [2,15,37] and GCN-based methods [6,7,16,21,25,34].

Table 3. Recognition accuracy comparison against state-of-the-art methods on the NTU RGB+D dataset.

Methods	X-Sub(%)	X-View(%)
Deep LSTM [24]	60.7	67.3
Ind-RNN [17]	81.8	88.0
TCN [2]	74.3	83.1
HCN [15]	86.5	91.1
SGN [37]	89.0	94.5
ST-GCN [34]	81.5	88.3
AS-GCN [16]	86.8	94.2
2 s-AGCN [25]	88.5	95.1
PT-GCN [38]	90.7	96.0
Shift-GCN [7]	90.7	96.5
MS-G3D [21]	91.5	96.2
MST-GCN [6]	91.5	**96.6**
CPT-GCN (Bone)	90.1	94.5
CPT-GCN (Joint+Bone)	91.9	96.2
CPT-GCN	**92.2**	96.5

Table 4. Recognition accuracy comparison against state-of-the-art methods on the NTU RGB+D 120 dataset.

Methods	X-Sub(%)	X-Set(%)
ST-LSTM [20]	55.7	57.9
SGN [37]	79.2	81.5
ST-GCN [34]	70.7	73.2
AS-GCN [16]	77.9	78.5
ST-Transformer [23]	82.7	84.7
2 s-AGCN [25]	82.9	84.9
PT-GCN [38]	85.0	87.3
Shift-GCN [7]	85.9	87.6
MS-G3D [21]	86.9	88.4
MST-GCN [6]	87.5	88.8
CPT-GCN	**88.9**	**89.8**

Our model achieves significant improvements of +1.4% and +1.0% over MST-GCN on the X-Sub and X-Set benchmark of NTU-120, respectively. Overall, CPT-GCN achieves better performance than other methods on both datasets, which demonstrates the superiority of our model.

5 Conclusion

In this work, we present a novel combined part-wise topology graph convolutional network (CPT-GCN) for skeleton-based action recognition. SPT-GC accurately learns the joint correlation of actions in a way that combines global topology and local topology. TPT-GC reasonably changes the receptive field of time convolution to extract the motion trend and motion details of the whole and part of the action. STPT-GC focuses on extracting the implicit spatio-temporal association information in the skeleton sequence, and establishes the part-enhanced spatio-temporal association topology. The combination of the three modules shows a powerful correlation modeling capability. We evaluate the proposed model on two large-scale datasets. The experimental results demonstrate that CPT-GCN has stronger performance than other graph convolutions, and the final model has excellent performance and generalization ability.

Acknowledgement. This work is supported by the Key Research and Development Program of China (No. 2022YFC3005401), the Key Research and Development Program of China, Yunnan Province (No. 202203AA080009), the Fundamental Research Funds for the Central Universities (No. B230205027), Postgraduate Research & Practice Innovation Program of Jiangsu Province(No. 422003261), the 14th Five-Year Plan for Educational Science of Jiangsu Province (No. D/2021/01/39), the Jiangsu Higher Education Reform Research Project (No. 2021JSJG143) and the 2022 Undergraduate Practice Teaching Reform Research Project of Hohai University.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. ACM Comput. Surv. (CSUR) **43**(3), 1–43 (2011)
2. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
4. Chen, T., et al.: Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4334–4342 (2021)
5. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13359–13368 (2021)

6. Chen, Z., Li, S., Yang, B., Li, Q., Liu, H.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1113–1122 (2021)
7. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 180–189 (2020)
8. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3288–3297 (2017)
11. Kim, T.S., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1623–1631. IEEE (2017)
12. Li, C., Huang, Q., Li, X., Wu, Q.: Human action recognition based on multi-scale feature maps from depth video sequences. Multimedia Tools Appl. **80**, 32111–32130 (2021)
13. Li, C., Huang, Q., Li, X., Wu, Q.: A multi-scale human action recognition method based on laplacian pyramid depth motion images. In: Proceedings of the 2nd ACM International Conference on Multimedia in Asia, pp. 1–6 (2021)
14. Li, C., Huang, Q., Mao, Y.: DD-GCN: directed diffusion graph convolutional network for skeleton-based human action recognition. In: IEEE International Conference on Multimedia and Expo (ICME) (2023)
15. Li, C., Zhong, Q., Xie, D., Pu, S.: Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 786–792 (2018)
16. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3595–3603 (2019)
17. Li, S., Li, W., Cook, C., Zhu, C., Gao, Y.: Independently recurrent neural network (INDRNN): building a longer and deeper RNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5457–5466 (2018)
18. Li, W., Wen, L., Chang, M.C., Nam Lim, S., Lyu, S.: Adaptive RNN tree for large-scale human action recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1444–1452 (2017)
19. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: NTU RGB+d 120: a large-scale benchmark for 3d human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2684–2701 (2019)
20. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3d human action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 816–833. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_50
21. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 143–152 (2020)

22. Peng, W., Hong, X., Chen, H., Zhao, G.: Learning graph convolutional network for skeleton-based human action recognition by neural searching. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 2669–2676 (2020)
23. Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12663, pp. 694–701. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68796-0_50
24. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+ d: a large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
25. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12026–12035 (2019)
26. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Trans. Image Process. **29**, 9532–9545 (2020)
27. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Action recognition via pose-based graph convolutional networks with intermediate dense supervision. Pattern Recogn. **121**, 108170 (2022)
28. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In: proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1227–1236 (2019)
29. Song, Y.F., Zhang, Z., Shan, C., Wang, L.: Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In: proceedings of the 28th ACM International Conference on Multimedia, pp. 1625–1633 (2020)
30. Thakkar, K., Narayanan, P.J.: Part-based graph convolutional network for action recognition. In: 29th British Machine Vision Conference, BMVC. p. Amazon et al. Microsoft; NVIDIA; SCANs; SCAPE. BMVA Press (2019)
31. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7912–7921 (2019)
32. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In: Proceedings of the Asian Conference on Computer Vision (2020)
33. Wang, W., Xie, W., Tu, Z., Li, W., Jin, L.: Multi-part adaptive graph convolutional network for skeleton-based action recognition. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2022)
34. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
35. Zhang, H., et al.: Resnest: split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2736–2746 (2022)
36. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2117–2126 (2017)

37. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1112–1121 (2020)
38. Zhu, X., Huang, Q., Li, C., Wang, L., Miao, Z.: Part-wise topology graph convolutional network for skeleton-based action recognition. In: Fang, L., Povey, D., Zhai, G., Mei, T., Wang, R. (eds.) Artificial Intelligence. CICAI 2022. LNCS, vol. 13604. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20497-5_26