Abstract. The topological structure of skeleton data plays a significant role in human action recognition. Combining the topological structure with graph convolutional networks has achieved remarkable performance. In existing methods, modeling the topological structure of skeleton data only considered the connections between the joints and bones, and directly use physical information. However, there exists an unknown problem to investigate the key joints, bones and body parts in every human action. In this paper, we propose the centrality graph convolutional networks to uncover the overlooked topological information, and best take advantage of the information to distinguish key joints, bones, and body parts. A novel centrality graph convolutional network firstly highlights the effects of the key joints and bones to bring a definite improvement. Besides, the topological information of the skeleton sequence is explored and combined to further enhance the performance in a four-channel framework. Moreover, the reconstructed graph is implemented by the adaptive methods on the training process, which further yields improvements. Our model is validated by two large-scale datasets, NTU-30007v1 and Kinetics, and outperforms the state-of-the-art methods.

Keywords: Centrality, Graph convolutional network, Skeleton data, Human action recognition.

1 Introduction

Human action recognition has attracted substantial attention from different research areas in recent years, since the deluge of skeleton data offers unprecedented opportunities to investigate structures, appearance, joints and bones, video sequences of human actions. The study of human action recognition provides significant insights into action surveillance, pose tracking, medical
monitor and diagnosis, sports analysis, security reports, and human-computer interaction [51, 49, 1, 17, 8, 2, 7]. Compared to traditional RGB or RGB-D video data, the skeleton sequences are compact motion data, which can significantly reduce computational cost on human action recognition. Besides, the skeleton data could show better performance when the background of human action is complicated [20, 56, 54].

The study of human action recognition, like walking, sitting and eating actions, has witnessed significant progress recently, like modeling the structures of human joints and analyzing their dynamic features. To accurately classify and predict human actions, many learning methods, such as recurrent neural networks (RNN) or convolutional neural networks (CNN), graph convolutional networks (GCN), and spatial-temporal graph convolutional networks (ST-GCN), have been proposed [15, 25, 56]. The earliest models for human action recognition often construct all coordinates of human joints from each frame for convolutional learning, like RNN or CNN [15, 25]. These learning models rarely investigate the endogenous relationships between joints and bones, leading to missing abundant unique and significant information. To emphasize the relationships between joints and bones, recent models build a skeleton graph, where nodes are human joints and edges are natural bones between two connected joints, and implement GCN to extract correlated features [24]. Later, the spatial-temporal GCN (ST-GCN) further developed GCN to simultaneously learn spatial features of skeleton sequences [56]. The ST-GCN constructed a graph to represent the connections of joints, and firstly proposed the temporal edges to link the relationships between consecutive frames. Although the ST-GCN captures the features of joints and bones, structurally key joints and bones are largely ignored, which may contain significant patterns and information of actions. For example, human hand joints and knee joints play key roles in the walking action. While the ST-GCN tries to capture all of the human joints with hierarchical GCN, the correlations between key joints and actions might weaken during the training process.

To address such issues, we need a new solution that can automatically detect the key human joints and bones, and deeply emphasize their relationships in terms of their large distance for every human action. Meanwhile, the new method could encode the highlighted features into the topological structures of the human skeleton as well as their dynamic properties, which could strongly elaborate on the strength of GCN, that is, highly clustering performance. While the skeletons are the structural graphs instead of 2D or 3D grids on different human actions, most of the previous models construct a constant graph as the input signal. Basically, to complete an action, everyone needs the cooperation of various body parts. The phenomenon indicates human actions are based on the relationships between different body parts, but some body parts may play key roles in human actions. For example, walking action strongly depends on legs, knees, and feet, while sitting action is closely related to the head and trunk. An exogenous dependency has been proposed for disconnected joints or human parts, but they could not cover wider-range human actions [59].
In this paper, we systematically study human actions, and propose a new and generic graph convolutional network to model characteristic skeletons for human action recognition, called the Centrality Graph Convolutional Networks (CGCN). In graph theory, centrality is defined to identify the most important nodes, edges or subgraphs within a graph. Centrality concepts were first developed in social network, and many of the terms used to measure centrality reflect their influential agent [38]. Applications include identifying the most influential persons in a social network, network flows, walk structure, and key infrastructure subgraphs in Internet networks [40,39,57]. Considering the skeleton data, this new model answers what characterizes important joints, bones and endogenous body parts for different actions. Then, the CGCN is expected to provide a ranking that identifies the most important joints, bones and body parts.

The CGCN provides new and significant insights for skeleton-based action recognition by highlighting the key joints, bones and body parts. As illustrated in Fig. 1, this graph model is formulated on top of a sequence of skeleton graphs, where each node reflects a joint on the human body. The basic structure learns the centrality features, like joint centrality, bone centrality and subgraph centrality, called the centrality module. Then, the extracted features are encoded into the spatial-temporal module for further training. The CGCN can model the relationships between physically connected and disconnected joints simultaneously, which could effectively capture low-level and high-order skeleton information. Besides, the centrality module via closeness, eigenvector and triplet subgraph calculations can obtain characteristic topological features from skeleton data. This module also reflects endogenous dependencies in an attention mechanism to enhance clustering performance.

Based on the intrinsic graph structures and human actions, we study new strategies to design centrality graph convolution networks, with inspirations from graph theory. The major contributions of this work are summarized in four aspects: (1) The CGCN is the first work for highlighting the centrality structures, like the key joints, bones and body parts on human actions. And it is designed to uncover the overlooked information between physically connected and disconnected parts of the human skeleton. (2) The CGCN follows several graph mechanisms in designing the centrality module to meet specific demands in human actions. It could offer a new and deep understanding of the action recognition task. (3) The motion information between consecutive frames is extracted for temporal information modeling. Both the spatial and motion information is fed into a four-channel framework for the action recognition task. (4) Our model outperforms the state-of-the-art methods on two large-scale datasets for skeleton-based action recognition. The find indicates that these centrality structures are fundamental mechanisms and factors hidden in the skeleton topology, which brings remarkable improvements for human action recognition.
2 Related work

**Skeleton-based action recognition.** Despite the illumination change and scene variation, the reliable skeleton data could be easily and accurately extracted by pose estimation algorithms from depth sensors [46] or RGB cameras [18, 55] and temporal CNN [31]. The substantial skeleton data spurs the creation of new technological or theoretical action recognition models [6, 48, 52]. Recently, deep learning approaches have been widely applied to build the spatial-temporal frameworks of skeleton sequences in human action recognition [42, 36]. For example, RNN models are implemented into human action recognition due to effectively learning long-term sequence data [32, 45]. And Jain combines a spatial-temporal graph with RNN to model the human body components, like arm, leg, and trunk, which finds a good way to map joints into a graph [19]. Meanwhile, compared to RNN, CNN also showed important implications for human action recognition, owing to its powerful parallelization over every element in the training process. Skeleton data are manually transferred into 2D or 3D images deployed by CNN [33], which achieves high performance in action recognition. However, these images used by CNN cannot fully reflect the topological structures of skeleton data. To address the limitations, GCN successfully applies to human action recognition, and its plausible mechanisms show more promising results [56]. The traditional GCN models only consider physical connections between different joints on human actions. Since the endogenous factors, referring to physically disconnected joints, also play a preeminent role in human action recognition, the skeleton structure conveys much unrevealed information and leaves potential space to improve the performance of GCN [53]. In this work,
we investigate key topological properties of the skeleton data, such as closeness, eigenvalue, and triplet subgraphs. By using the centrality module, we successfully encode these characteristic features into the GCN framework and highlight the key joints, bones, and body parts for learning better human actions.

**Graph convolutional networks.** Mapping the relational data into graphs, the topological structures can be encoded to model the connections among nodes, and provide more promising perspectives underlying the data. Inspiring by this mechanism, GCN is successfully implemented in deep learning research. The various approaches in modeling GCNs fall into two categories, including spatial and spectral approaches. Spatial approaches use graph theory to define the nodes and edges for entities on data [12,3]. Interestingly, spectral approaches analyze the constructed graph in the frequency domain [16,9]. The spectral approach usually leverages the Laplacian eigenvector to transform a graph in the time domain to in frequency domain, potentially resulting in large computation cost [16]. Considering human action recognition, most of the methods choose the spatial approaches to construct the GCN due to the large size of skeleton data. Since the human body is naturally formed as a graph, not a sequence or an image, the related features are easily extracted from skeleton data. However, their works only focus on the static graph structure and are hard to understand the dynamic information from human actions. Here, our model could adaptively extract the key information on joints, bones and body parts, which yields a dynamical GCN. Besides, the CGCN combines the high-order information to provide a new insight for learning human actions.

3 Method

In this section, we explicate the centrality GCN for human action recognition based on skeleton data. Firstly, we introduce how to build the topological structures of a spatial graph from skeleton data.

3.1 Spatial graph construction

A graph is a diagrammatical representation of the skeleton structure. It consists of nodes and edges and can be defined to $G = (N, E)$, where $N$ is the number of joints, and $E$ is the number of bones. $A$ is considered as its adjacency matrix with $N \times N$ dimensions. The entry $A(i, j)$ is 1 if the edge connects node $i$ to node $j$; or 0 otherwise. Let $x \in \mathbb{R}^N$ be a feature vector for every node in a graph. By Fourier transformation, a spectral filter is defined to $g(\theta) = \text{diag}(\theta)$, where $\theta \in \mathbb{R}^N$ is a parameter vector. Thus, the spectral convolution on a graph can be formulated as

$$z = g(\theta) \star x = U g(\theta) U^T x,$$

where $z$ is the extracted feature vector for each node, and $U$ is the orthogonal matrix of eigenvectors of the Laplacian matrix $L$, i.e. $L = U \Lambda U^T =$
\( I_{N \times N} - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \) (\( D_i = \sum_j A(i, j) \) is the degree vector). Through the Laplacian transform, the \( g(\theta) \) can be considered as a function of the eigenvalues of \( L \), i.e. \( g(\theta)(A) \). Since the complexity time of \( UAU^T \) is \( O(N^2) \), solving Eq. 1 is computationally expensive. In order to reduce time cost for large graphs, \( g(\theta)(A) \) can be approximated by a truncated expansion by Chebyshev polynomials \( T_m(x) \) up to \( M^{th} \) order \[56\]

\[
g(\theta')(A) \approx \sum_{m=0}^{M} \theta'_m T_m(\hat{\Lambda}), \tag{2}
\]

where \( \hat{\Lambda} = \frac{2}{\lambda_{max}} L - I_{N \times N} \), and \( \theta' \in R^M \) is a vector of Chebyshev coefficients. \( \lambda_{max} \) is the maximum eigenvalue of \( L \). Generally speaking, the Chebyshev polynomials are two sequences of polynomials, and defined by the recurrence relation: \( T_m(x) = 2xT_{m-1}(x) - T_{m-2}(x) \), with \( T_0(x) = 1 \) and \( T_1(x) = x \).

Thus, with the approximating method, the Eq. 1 can be addressed

\[
z = g(\theta') \ast x \approx \sum_{m=0}^{M} \theta'_m T_m(\hat{\Lambda})x, \tag{3}
\]

where \( \hat{\Lambda} = \frac{2}{\lambda_{max}} L - I_{N \times N} \), and \((UAU^T)^m = UA^mU^T\). Note that this formula is \( M \)-localized since it is a \( M^{th} \)-order polynomial in the Laplacian matrix, that is, it is determined by nodes that are at maximum \( M \) steps away from the central node (\( M^{th} \)-order neighborhood). The complexity time of Eq. 3 is \( O(E) \), where \( E \) is the number of edges. Defferrard et al. (2016) use this \( M \)-localized convolution to successfully implement a convolutional network on graphs.

In order to obtain the linear Laplacian spectrum, we set \( M = 1 \) and \( \lambda_{max} = 2 \) in Eq. 3. By this method, we could derive a linear formulation of Eq. 3 as

\[
z = g(\theta') \ast x \approx \theta'_0 x + \theta'_1 (L - I_{N \times N})x = \theta'_0 x - \theta'_1 D^{-\frac{1}{2}}AD^{-\frac{1}{2}}x, \tag{4}
\]

with two parameters \( \theta'_0 \) and \( \theta'_1 \). Here, Eq. 4 allows us to construct deeper models to improve their capacity on large graphs. Besides, the two parameters could adapt to the change in scale during training.

In practice, minimizing the number of operations per layer, and Eq. 4 can be simplified to

\[
z = g(\theta) \ast x \approx \theta \left( I_{N \times N} + D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \right)x, \tag{5}
\]

where \( \theta = \theta'_0 = -\theta'_1 \), and the eigenvalues of \( I_{N \times N} + D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \) range from 0 to 2.

For Eq. 5, the normalized formulation can be written as

\[
Z = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}X\Theta
= \tilde{A}X\Theta, \tag{6}
\]
where $\tilde{A} = \hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} = I_{N \times N} + D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, $\hat{D}_t = \sum_j \hat{A}_{ij}$, $X \in R^{N \times C}$ and $\Theta \in R^{C \times F}$ ($C$ is the dimensions of feature vector per node and $F$ is the filter channels). Note that $Z$ is the convolved or extracted feature matrix, and $\Theta$ is the filtering-parameters matrix. Thus, the complexity of this normalized formulation is $O(EFC)$.

### 3.2 Centrality graph convolutional networks

Considering human actions, the skeleton graphs are the sequence graphs or temporal graphs, that is, $G = G_1, G_2, G_3, ..., G_t$, where $t$ is the time steps. Besides, human action recognition can be defined as a supervised-learning classification problem. Based on GCN, the proposed model, centrality graph convolutional networks, consists of three modules, such as skeleton-data input module, centrality module, Softmax module. The schematic of CGCN is illustrated in Fig. 1. In the first module, the joints and bones of the skeleton are extracted from the input images, i.e., $X_{in}$. The last module obtain the trained features and model (i.e., $X_{out}$) to calculate the action scores, and then make a prediction [56]. The function of the centrality module plays a significant role in highlighting the key joints, bones and body parts during training. It has three components, namely, joint centrality, bone centrality, subgraph centrality. Next, the components of the centrality module will be explicated below.

**Joint centrality.** In the directed skeleton graph, the joint can be considered as a node in the graph. Meanwhile, every node has the corresponding coordinates. And different human action has different coordinates in the spatial domain, that is, the spatial distance between two nodes is changed on human action. In graph theory, the normalized closeness centrality is a reasonable solution to measure the relevance between two nodes in the spatial distance. In a connected graph, the closeness centrality of a node is the average length of the shortest path between the node and all other nodes [4]. It can be defined as the reciprocal of the distance

$$N_c(i) = \frac{N - 1}{\sum_j d(i, j)}, \quad (7)$$

where $d(i, j)$ is the spatial distance between node $i$ and $j$, $N$ is the number of nodes in the graph (normalizing the closeness value). Basically, the more central a node is, the closer it is to all other nodes. Measuring distances between a node and other nodes are irrelevant in undirected graphs, whereas it can yield highly correlate results in directed graphs. Indeed, our skeleton graph is a directed graph. Inspired by the mechanism, the skeleton graph can easily capture a high closeness centrality from outgoing edges, but low closeness centrality from incoming edges. This method can efficiently highlight the key joints (nodes) in the skeleton graphs for various human actions. To encode the significant and potential features into training, the closeness matrix of the skeleton graph is set as $J \in R^{N \times N}$. Note that $J$ is a symmetric matrix, has same size with the adjacency matrix.
**Bone centrality.** In the skeleton graph, the bone can be served as an edge. Due to the different coordinates of each node, the distances of bones have various changes. Considering graph theory, the edge betweenness centrality is a feasible method to measure the importance degree of an edge within a graph. The edge betweenness centrality is defined as the number of the shortest paths that pass through a specific edge in a graph or network [14]. Each edge in the skeleton graph can be associated with an edge betweenness centrality value. It can be written as

\[
B_c(e_{i,j}) = \sum_{(l,q) \neq (i,j)} \frac{\tilde{S}_{l,q}(e_{i,j}) d_{l,q}(e_{i,j})}{\tilde{S}_{l,q} d_{l,q}},
\]

where \( e_{i,j} \) is the edge between node \( i \) and node \( j \), \( \tilde{S}_{l,q} \) is the number of all existing shortest paths from node \( l \) to node \( q \), \( \tilde{S}_{l,q}(e_{i,j}) \) is the number of all shortest paths from node \( l \) to node \( q \) that pass through edge \( e_{i,j} \), \( d_{l,q} \) is the total distance length from node \( l \) to node \( q \), and \( d_{l,q}(e_{i,j}) \) is the total distance length from node \( l \) to node \( q \) that pass through edge \( e_{i,j} \). An edge with a high edge betweenness centrality score represents a bridge-like connection between two parts of a graph. If this edge is removed, it will directly affect the connectivity between many pairs of nodes through this edge. Maybe the removal of this edge leads to a partition of the skeleton graph into two densely connected subgraphs. Thus, the edge betweenness can directly highlight the importance of edges in the skeleton graph. The edge betweenness scores are also encoded into the betweenness matrix, i.e. \( B \in R^{N \times N} \).

**Subgraph centrality.** Different body parts are associated with complete human actions. However, there exists a problem of how to evaluate the influence of every body part in human actions and measure the correlations between two body parts. In graph theory, the body part could be seen as a subgraph. A subgraph \( S \) is a graph whose node set \( N(S) \) is a subset of the node set \( N(G) \), i.e. \( N(S) \subseteq N(G) \). Similarly, the subgraph edge set \( E(S) \) is a subset of the edge set \( E(G) \), that is \( E(S) \subseteq E(G) \). In 2016, Benson et al. [1] investigated the triplet subgraph in directed graphs, and proposed a feasible and efficient method to encode subgraph features into every corresponding node. Here, let \( S_1, S_2, S_3, ..., S_k \) denote a set of subgraphs (these subgraphs could have different node sets; in this work, we only consider 3-node subgraphs. For every node belonging to a subgraph \( S_k \), its weight will be counted by 1. Thus the subgraph-weighted matrix, named \( W \in R^{N \times N} \), is built by this method. The associations with many pairs of subgraphs are measured by every node weights, that is, a weighted adjacency matrix. Through the subgraph centrality, strong connections are built through disconnected body parts, which could highlight the key body parts of human actions.

### 3.3 Implementing CGCN

To implement the CGCN in human action recognition, let \( X = X_{in} \) be the input feature matrix on total frames per one video sample. Then, Eq.6 can be
rewritten as

\[ Z = \tilde{A}X\Theta \]
\[ = \tilde{A}X_{in}\Theta, \] (9)

Considering the proposed centrality matrixes \( J, B, W, \tilde{A} \) is replaced by \( J + B + W + \tilde{A} \). Solving Eq. (9) yields

\[ Z = (J + B + W + \tilde{A})X_{in}\Theta. \] (10)

Using Softmax classifier, one has

\[ Z = \text{Softmax}\left((J + B + W + \tilde{A})\text{ReLU}((J + B + W + \tilde{A})X\Theta^0)\Theta^1\right) \]
\[ = \text{Softmax}\left(\hat{C}\text{ReLU}(\hat{C}X\Theta^0)\Theta^1\right), \] (11)

where \( \hat{C} = J + B + W + \tilde{A} \), \( \Theta^0 \in R^{C \times N} \) is a weight matrix with \( N \) feature maps, and \( \Theta^0 \in R^{N \times F} \) is a hidden-to-output weight matrix. The softmax operation normalizes the combinations of four matrixes.

3.4 Network architecture

One convolutional block for the CGCN is based on ST-GCN, i.e., implementing spatial and temporal convolution on three dimensions of feature maps, including the joint features \( (C) \), the frame number \( (T) \), the number of joints \( (N) \). In Fig. 2, one basic block of the CGCN is composed of seven layers, namely a spatial convolution (S-conv) layer, two batch normalization (BN) layers, two ReLU layers, a temporal convolution (T-conv) layer and a dropout layer. Note that the drop rate is set to 0.5. Note that the spatial convolution layer contains four different streams, such as joint centrality stream, bone centrality stream, subgraph centrality stream, and adjacency stream.

The centrality graph convolutional network (CGCN) has nine basic convolution blocks. The numbers of output channels are 64, 64, 64, 128, 128, 128, 256, 256 and 256 on every block. At the beginning of nine basic blocks, there exists a data BN layer, which normalizes the input skeleton data. Through nine basic blocks, it is followed by a global average pooling layer, adjusting different feature dimensions of video samples into the same dimensions. Finally, the final results are calculated by a Softmax classifier layer to yield the classification results.

According to the architecture of CGCN in Fig. 1, the extracted topology information, i.e., the joint centrality, bone centrality and subgraph centrality, plays a significant role for human action recognition but is ignored in previous methods. Thus, we develop a new method to extract topology information with a four-channel framework to improve recognition.

In the directed skeleton graph, each bone is linked by two joints, such as the source node and the target node, respectively. The source node is close to the gravity point of the skeleton, while the target node is far away from that. A
directed edge is defined as a bone between the source node and the target node. Note that every directed edge has the length of the corresponding bone and the direction information. For example, given two joints coordinates, that is, the source node \( n_i = (x_i, y_i, z_i) \) and the target node \( n_j = (x_j, y_j, z_j) \), the length of the bone can be calculated as 
\[
d(n_i, n_j) = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}.
\]
In this directed skeleton graph, there is no cycle, which indicates every directed edge can only contain a source node and a target node. In this way, the skeleton adjacency, joint centrality, bone centrality and subgraph centrality are calculated and implemented into graph convolution blocks, respectively. Finally, the Softmax classifier calculates the results of four streams to yield the score ranking and predict the action classification.

4 Experiment

To evaluate the proposed CGCN model, we conduct the experiments on two large-scale action recognition datasets: NTU-RGB+D [43,35] and Kinetics [21]. First, since the size of the NTU-RGB+D dataset is smaller than that of the Kinetics dataset, we perform the ablation studies of our model on the NTU-RGB+D dataset, which examines every component of the proposed model on the action recognition performance. Then, the performance of the CGCN model is verified and compared with that of the other state-of-the-art approaches.

4.1 Datasets

**NTU-RGB+D:** NTU-RGB+D is currently one of the largest dataset widely used in the skeleton-based action recognition, which consists of 56,880 video samples categorized into 60 action classes. All of video samples are acted by 40 volunteers, who come from different age groups from 10 to 35. Each action is recorded by three Kinect V2 cameras concurrently from different horizontal
angles, like $-45^\circ, 0^\circ, 45^\circ$. This dataset contains 3D skeletal data for each video sample. Here, it records 25 joints for each frame on each subject sample, while each video sample has one or two subjects. The dataset recommends two benchmarks. The first one is the cross-subject (X-sub), which is composed of a training set (40,320 video samples) and a testing set (16,560 video samples). The second one is the cross-view (X-view), where the training set contains 37,920 video samples captured by cameras No. 2 and No. 3, and the testing set contains 18,960 video samples captured by camera No. 1. In the comparison, we report the top-1 accuracy on two benchmarks.

**Kinetics:** Kinetics is a large-scale dataset for human action recognition, containing 300,000 video samples. All of video samples covers 400 human action classes. The video samples are recorded by YouTube and have various subjects. It only contains RGB video samples without raw skeleton data. We obtain 18 joints’ positions of every frame using the publicly available OpenPose toolbox [5]. The captured skeleton data contains two dimensions of coordinates, i.e., $(x,y)$, and confidence score $(c)$ for each joint. The dataset consists of a training set (240,000 video samples) and a testing set (20,000 video samples). Following the evaluation method, we report the top-1 and top-5 accuracies on the benchmark.

### 4.2 Training details

The CGCN model is implemented on the PyTorch deep learning framework [41]. The kernel size of this model is set to 4. And it contains four channels to train the skeleton data. The optimization method uses the stochastic gradient descent with Nesterov momentum (0.9). And the loss function chooses the cross-entropy for backpropagating gradients. The batch size is 32 for 65 epochs. The CGCN model is trained on 2 GTX-1080Ti GPUs. For different datasets, we set the specific configurations of the CGCN model.

For the NTU-RGB+D dataset, each video sample contains no more than two persons. The feature dimensions are respectively 64, 128 and 256. The number of frames in each video sample is set to 300. If the frame number is less than 300 frames, we repeat the video sample until it reaches 300 frames. In Eq. (10) $X_{in}$ should be the total features from 300 frames. The learning rate is set to 0.1 and the decay rate is set to 0.0001. The training process ends at the 49th epoch.

For the Kinetics dataset, the input configuration of Kinetics is set to 150 frames with 2 persons in each video sample. We select the same data-augmentation methods as ST-GCN. For example, we randomly select 150 frames from the input skeleton data, and disorder the joint coordinates with randomly chosen translations and rotations. Here, $X_{in}$ should be the total features from 150 frames in Eq. (10). The learning rate is also set as 0.1 and the decay rate is set to 0.0001. The training process converges to the 57th epoch.
Table 1. Comparisons of the top-1 accuracy when adding centrality graph convolutional block with or without $J$, $B$ and $W$.

| Methods                     | Accuracy(%) |
|-----------------------------|-------------|
| ST-GCN without adaptive     | 88.3        |
| ST-GCN with adaptive        | 90.8        |
| CGCN without $J$            | 95.1        |
| CGCN without $B$            | 95.9        |
| CGCN without $W$            | 95.9        |
| CGCN                        | 96.4        |

4.3 Ablation study

We examine the influence of each centrality components in the centrality graph convolutional network (CGCN) with the X-view benchmark on the NTU-RGB+D dataset. The performance of ST-GCN on the NTU-RGB+D dataset is 88.3%. By highlighting the centrality information and the specially four-channel training framework, the result is improved to 96.4%, which brings a definite improvement. The detail is introduced in the below parts.

Centrality graph convolutional block. As mentioned in Section 3.2, there are four types of centrality structures in the centrality graph convolutional block, i.e., $A$, $J$, $B$, and $W$. Note that $A$ is the original skeleton graph, namely, adjacency matrix. Note that $A$ is the same as the adjacency matrix in ST-GCN. Here, we only investigate the effects of the other three centrality structures. The different combinations of the centrality structures are tested on CGCN, including CGCN without $J$, CGCN without $B$ and CGCN without $W$, and the results are shown in Table 1. The results show that each centrality structure extracted from the graph is significant for human action recognition. Besides, deleting any one of centrality structures will dramatically reduce the performance. After combining all centrality structures added together, we can utilize all in information in each centrality structure and achieve better outcomes. We also test the influence of the adaptive skeleton graph used in the original ST-GCN. The result indicates that the weight and dynamic adjacency matrix is important, which also demonstrates the significance of the adaptive skeleton graph.

Visualization of the centrality graphs. Fig. 3 is a visualization of the three centrality graph for one layer on the same action (from left to right is the joint centrality, bone centrality, and subgraph centrality, respectively). The skeleton graphs are plotted from the physical connections of the human body. The red joints, bones, body parts represent the highlighted parts by the centrality graphs. In Fig. 3, it indicates that the joints and bones of the human trunk play significant roles in this action. Besides, it also suggests that a traditional skeleton graph is not the best choice for the action recognition task, and different actions need
Fig. 3. Illustration of the three centrality graphs for one layer on the same action. Red joints, bones and body parts are highlighted by the centrality graphs, namely, key joints, bones and body parts.

graphs with different centrality structures. The joint centrality graph pays more attention to the adjacent joints in the physical skeleton graph. For the subgraph centrality structures, the neck part and the hipbone part are detected to have a stronger connection, although they are far away from each other in the physical skeleton. Compared to previous works, the centrality graphs further capture the low-level features, like key joints and bones, and high-order structures, namely body parts. This information could uncover endogenous factors on different human actions and give new perspectives on human action recognition and graph convolutional networks. Thus, the centrality structures are more relevant to the action classification task.

4.4 Comparison with the state-of-the-art

We compare the performance of our model with the state-of-the-art action recognition methods on the NTU-RGB+D dataset [11, 43, 34, 47, 58, 28, 30, 23, 22, 37, 56, 10, 26, 50, 44] and Kinetics dataset [13, 43, 23, 56, 29, 44]. The results of these two comparisons are shown in Table 2 and Table 3 respectively. The methods used for comparison are categorized as RNN-based methods, CNN-based methods, and GCN-based methods. Importantly, our model outperforms the state-of-the-art methods on both datasets, which validate the superiority of our model.

5 Conclusions

In this work, we propose a novel centrality graph convolutional network (CGGCN) for human action recognition. It captures the key joints, bones and body parts from human actions and embeds them into the graph convolution networks to adaptively learn and update. Besides, the existing methods always ignore or overlook the importance of centrality information on skeleton data, i.e., joint centrality, bone centrality, and subgraph centrality. However, these centrality structures could reflect the low-level features and high-order information of the skeleton data. Moreover, a four-channel framework explicitly employs these centrality
Table 2. Comparisons of the top-1 accuracy with state-of-the-art methods on the NTU-RGB+D dataset

| Methods               | X-sub (%) | X-view (%) |
|-----------------------|-----------|------------|
| H-RNN [11]            | 59.1      | 64.0       |
| Deep LSTM [43]        | 60.7      | 67.3       |
| 2s-3DCNN [34]         | 66.8      | 72.6       |
| ST-LSTM [36]          | 69.2      | 77.7       |
| STA-LSTM [47]         | 73.4      | 81.2       |
| VA-LSTM [58]          | 79.2      | 87.7       |
| ARRN-LSTM [28]        | 80.7      | 88.8       |
| Ind-RNN [30]          | 81.8      | 88.0       |
| TCN [23]              | 74.3      | 83.1       |
| C-CNN+MTLN [22]       | 79.6      | 84.8       |
| Synthesized CNN [37]  | 80.0      | 87.2       |
| ST-GCN [50]           | 81.5      | 88.3       |
| CNN-Motion+Trans [10] | 83.2      | 89.3       |
| ResNet152 [26]        | 85.0      | 92.3       |
| DPRL+GCNN [50]        | 83.5      | 89.8       |
| 2s-AGCN [44]          | 88.5      | 95.1       |
| CGCN (ours)           | 90.3      | 96.4       |

Table 3. Comparisons of the top-1 and top-5 accuracies with state-of-the-art methods on the Kinetics-Skeleton dataset.

| Methods               | Top-1 (%) | Top-5 (%) |
|-----------------------|-----------|-----------|
| Feature Encoding [13] | 14.9      | 25.8      |
| Deep LSTM [43]        | 16.4      | 35.3      |
| TCN [23]              | 20.3      | 40.0      |
| ST-GCN [50]           | 30.7      | 52.8      |
| AS-GCN [29]           | 34.8      | 56.3      |
| 2s-AGCN [44]          | 36.1      | 58.7      |
| CGCN (ours)           | 37.5      | 60.4      |

structures, which further improves performance. Finally, the CGCN model is validated on two large-scale datasets, namely NTU-RGB+D and Kinetics, and it outperforms the state-of-the-art methods on two datasets. The finding indicates that these centrality structures are fundamental mechanisms and factors hidden in the skeleton topology. This novel model sheds some new lights on future studies of skeleton-based action recognition. For example, how to incorporate the correlations between adjacent frames, such as similarities, distances, and structures into CGCN becomes a natural question.
References

1. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5167–5176 (2018)

2. Athalye, A., Carlini, N.: On the robustness of the cvpr 2018 white-box adversarial example defenses. arXiv preprint arXiv:1804.03286 (2018)

3. Atwood, J., Towsley, D.: Diffusion-convolutional neural networks. In: Advances in neural information processing systems. pp. 1993–2001 (2016)

4. Bavelas, A.: Communication patterns in task-oriented groups. The journal of the acoustical society of America 22(6), 725–730 (1950)

5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7291–7299 (2017)

6. 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)

7. Chakraborty, B.K., Sarma, D., Bhuyan, M.K., MacDorman, K.F.: Review of constraints on vision-based gesture recognition for human–computer interaction. IET Computer Vision 12(1), 3–15 (2017)

8. Chen, J., Little, J.J.: Sports camera calibration via synthetic data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)

9. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in neural information processing systems. pp. 3844–3852 (2016)

10. Du, Y., Fu, Y., Wang, L.: Skeleton based action recognition with convolutional neural network. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). pp. 579–583. IEEE (2015)

11. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1110–1118 (2015)

12. Duvenaud, D.K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., Adams, R.P.: Convolutional networks on graphs for learning molecular fingerprints. In: Advances in neural information processing systems. pp. 2224–2232 (2015)

13. Fernando, B., Gavves, E., Oramas, J.M., Ghodrati, A., Tuytelaars, T.: Modeling video evolution for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5378–5387 (2015)

14. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the national academy of sciences 99(12), 7821–7826 (2002)

15. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp. 6645–6649. IEEE (2013)

16. Henaff, M., Bruna, J., LeCun, Y.: Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163 (2015)

17. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J.: Artificial intelligence in radiology. Nature Reviews Cancer 18(8), 500–510 (2018)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1125–1134 (2017)
19. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 5308–5317 (2016)
20. Jiang, Y., Upputuri, P.K., Xie, C., Lyu, Y., Zhang, L., Xiong, Q., Pramanik, M., Pu, K.: Broadband absorbing semiconducting polymer nanoparticles for photoacoustic imaging in second near-infrared window. Nano letters 17(8), 4964–4969 (2017)
21. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
22. 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)
23. 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)
24. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
26. Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., He, M.: Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 601–604. IEEE (2017)
27. Li, C., Zhong, Q., Xie, D., Pu, S.: Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. arXiv preprint arXiv:1804.06055 (2018)
28. Li, L., Zheng, W., Zhang, Z., Huang, Y., Wang, L.: Skeleton-based relational modeling for action recognition. arXiv preprint arXiv:1805.02556 1(2), 3 (2018)
29. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
30. 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)
31. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. arXiv preprint arXiv:1705.04724 (2017)
32. 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)
33. Li, Y., Wang, N., Liu, J., Hou, X.: Demystifying neural style transfer. arXiv preprint arXiv:1701.01036 (2017)
34. Liu, H., Tu, J., Liu, M.: Two-stream 3d convolutional neural network for skeleton-based action recognition. arXiv preprint arXiv:1705.08106 (2017)
35. 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.
36. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: European conference on computer vision. pp. 816–833. Springer (2016)
37. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition **68**, 346–362 (2017)
38. Newman, M.: Networks. Oxford university press (2018)
39. Newman, M.E.: A measure of betweenness centrality based on random walks. Social networks **27**(1), 39–54 (2005)
40. Okamoto, K., Chen, W., Li, X.Y.: Ranking of closeness centrality for large-scale social networks. In: International workshop on frontiers in algorithmics. pp. 186–195. Springer (2008)
41. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
42. Rahmani, H., Mian, A., Shah, M.: Learning a deep model for human action recognition from novel viewpoints. IEEE transactions on pattern analysis and machine intelligence **40**(3), 667–681 (2017)
43. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
44. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12026–12035 (2019)
45. Shi, Z., Kim, T.K.: Learning and refining of privileged information-based rnns for action recognition from depth sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3461–3470 (2017)
46. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1297–1304. Ieee (2011)
47. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Thirty-first AAAI conference on artificial intelligence (2017)
48. Stergiou, A., Poppe, R., Veltkamp, R.C.: Learning class regularized features for action recognition. arXiv preprint arXiv:2002.02651 (2020)
49. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6479–6488 (2018)
50. Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J.: Deep progressive reinforcement learning for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5323–5332 (2018)
51. Tuttle, R.M., Fagin, J.A., Minkowitz, G., Wong, R.J., Roman, B., Patel, S., Untch, B., Ganly, I., Shaha, A.R., Shah, J.P., et al.: Natural history and tumor volume kinetics of papillary thyroid cancers during active surveillance. JAMA Otolaryngology–Head & Neck Surgery **143**(10), 1015–1020 (2017)
52. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
53. Weng, J., Weng, C., Yuan, J.: 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. pp. 4171–4180 (2017)
54. Xie, C., Li, C., Zhang, B., Chen, C., Han, J., Zou, C., Liu, J.: Memory attention networks for skeleton-based action recognition. arXiv preprint arXiv:1804.08254 (2018)
55. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose flow: Efficient online pose tracking. arXiv preprint arXiv:1802.00977 (2018)
56. 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)
57. You, K., Tempo, R., Qiu, L.: Distributed algorithms for computation of centrality measures in complex networks. IEEE Transactions on Automatic Control 62(5), 2080–2094 (2016)
58. 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)
59. Zhang, P., Lan, C., Zeng, W., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. arXiv preprint arXiv:1904.01189 (2019)