Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition

Zhan Chen,[1] Sicheng Li,[2] Bing Yang,[1] Qinghan Li,[3] Hong Liu[1]*

[1]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, China
[2]Institute of Information and Communication Engineering, Zhejiang University, China
[3]Department of Electrical Engineering and Computer Science, Syracuse University, USA
zhanchen_cz@pku.edu.cn, jasonlisicheng@zju.edu.cn, bingyang@sz.pku.edu.cn, Qli160@syr.edu, hongliu@pku.edu.cn

Abstract

Graph convolutional networks have been widely used for skeleton-based action recognition due to their excellent modeling ability of non-Euclidean data. As the graph convolution is a local operation, it can only utilize the short-range joint dependencies and short-term trajectory but fails to directly model the distant joints relations and long-range temporal information that are vital to distinguishing various actions. To solve this problem, we present a multi-scale spatial graph convolution (MS-GC) module and a multi-scale temporal graph convolution (MT-GC) module to enrich the receptive field of the model in spatial and temporal dimensions. Concretely, the MS-GC and MT-GC modules decompose the corresponding local graph convolution into a set of sub-graph convolution, forming a hierarchical residual architecture. Without introducing additional parameters, the features will be processed with a series of sub-graph convolutions, and each node could complete multiple spatial and temporal aggregations with its neighborhoods. The final equivalent receptive field is accordingly enlarged, which is capable of capturing both short- and long-range dependencies in spatial and temporal domains. By coupling these two modules as a basic block, we further propose a multi-scale spatial temporal graph convolutional network (MST-GCN), which stacks multiple blocks to learn effective motion representations for action recognition. The proposed MST-GCN achieves remarkable performance on three challenging benchmark datasets, NTU RGB+D, NTU-120 RGB+D and Kinetics-Skeleton, for skeleton-based action recognition.

Figure 1: Some samples from NTU RGB+D dataset. (a) "walking" (b) "hand waving" (c) "jumping up" (d) "cheering up" (e) "wearing on glasses" (f) "taking off glasses"

Introduction

Human action recognition has received a lot of attention in various applications, such as video surveillance, human-machine interaction, virtual reality, video analysis and so on (Poppe 2010; Weinland, Ronfard, and Boyer 2011; Aggarwal and Ryoo 2011; Sudha et al. 2017). The current studies on human action recognition are mainly divided into RGB video based and skeleton based. Compared with RGB video based methods which require high-cost computing resources to process pixel-level information in RGB image or temporal optical flow (Tran et al. 2015; Simonyan and Zisserman 20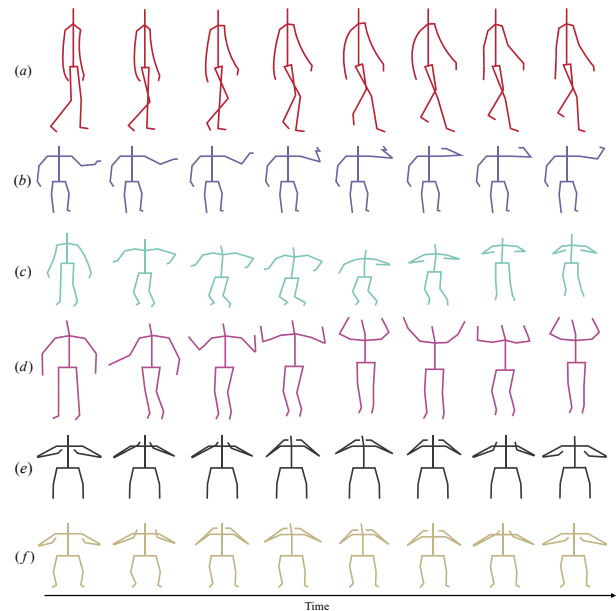14; Zhang et al. 2016; Li et al. 2020), skeleton based methods are more computationally efficient as skeleton data represents a human structure with 2D or 3D coordinates of a few dozen joints. Moreover, skeleton data conveys relatively high-level information for human actions, which exhibits stronger robustness to appearance variation and environmental noises, *e.g.*, background clutter, illumination changes (Wang et al. 2018; Vemulapalli, Arrate, and Chellappa 2014; Liu, Liu, and Chen 2017; Tu, Liu, and Liu 2018). Therefore, human action recognition with skeleton data has been attracting increasing attention.

To extract discriminative spatial and temporal features for skeleton-based action recognition, great efforts have been made to learn patterns embedded in the spatial configuration of the joints and their temporal dynamics. Conventional deep-learning-based methods learn human action representations with convolutional neural networks (CNNs) and

---

**Corresponding author

recurrent neural networks (RNNs). These methods manually construct the skeleton sequence as a kind of grid-shape structure such as a pseudo-image (Ke et al. 2017; Soo Kim and Reiter 2017; Li et al. 2017) or a sequence of the coordinate vectors (Du et al. 2015; Zhang et al. 2017; Liu et al. 2016). As the skeleton is naturally structured as a graph in non-Euclidean geometric space, these methods fail to fully explore the inherent relationships between human joints . Recently, Graph convolutional network (GCN) - based method (Yan, Xiong, and Lin 2018) addressed these drawbacks and constructed a spatial temporal skeleton graph with the joints as graph nodes and natural connectivities in both human body structures and time as graph edges, therefore, the spatial temporal relations between human joints are well embedded in the adjacency matrix of the skeleton graph.

The follow up works (Shi et al. 2019a,b; Si et al. 2019) have proposed many effective variants of GCN, they mostly consider the short-range connections. However, long-range dependencies are also important for action recognition. For instance, as shown in Fig. 1, (1) "Walking" requires coordination of the whole body to maintain balance, while "hand waving" can be accomplished just by hand. Since different actions require coordination of different body parts to complete, it is of critical importance to design a multi-scale feature extractor to capture the dependencies between different range joints. (2) "Jumping up" and "cheering up" could be rather ambiguous when only considering the local movements of the leg parts. Therefore, capturing the essential contextual information of an action which usually requires a large receptive field could help us to judge whether the jump is just a "jumping up" or a jump for "cheering up". (3) "Wearing on glasses" and "taking off glasses" are very ambiguous in a short period, which requires the algorithm to capture long-range temporal information and reduce ambiguity. In summary, capturing both short-range joints dependencies and distant joints relations in the spatial domain and considering both short-term trajectory and long-range temporal information are vital to skeleton-based action recognition. There are many GCN-based works considering one or both of those aspects (Li et al. 2019a; Shi et al. 2019c; Zhang, Xu, and Tao 2020; Peng et al. 2020; Liu et al. 2020; Li et al. 2019b). Nevertheless, the problem is far from being solved.

To capture both short-range joints dependencies and distant joints relations, we proposed a simple yet efficient multi-scale spatial graph convolution (MS-GC) module which is capable of extracting multi-scale spatial features. The MS-GC replaces the spatial graph convolution operation with a group of sub-graph convolutions which formulate a hierarchical structure with residual connections between adjacent subsets. When the features go through the module, they realize multiple information exchanges with neighboring nodes, and the equivalent spatial receptive field is thus increased to capture the relations between distant joints. For multi-scale temporal modeling, we naturally extend the MS-GC to the temporal domain and propose a multi-scale temporal graph convolution (MT-GC) module. With a similar architecture as the MS-GC module, the MT-GC module is

capable of modeling the long-range temporal relationship over distant frames. These two modules are complementary and cooperate in endowing the network with both multi-scale spatial and multi-scale temporal modeling abilities. We devise two approaches to combine these two modules as a basic block and propose a multi-scale spatial temporal graph convolution network (MST-GCN) by stacking multiple blocks. Experiments on three large-scale datasets for skeleton-based action recognition verify the effectiveness of the obtained model: when the parameters of MST-GCN are less than one third of ST-GCN, our model perform much better than ST-GCN, and when the model has parameters equivalent to ST-GCN, it achieves remarkable results compared to the state-of-the-art methods. Our contributions can be summarized as follows:

- A multi-scale spatial graph convolution module to capture both local joints connectives and non-local joints relations for spatial modeling.
- A multi-scale temporal graph convolution module to efficiently enlarge the temporal receptive field for long-range temporal dynamics.
- Integrating the MS-GC and MT-GC modules, we propose a multi-scale spatial temporal graph convolutional network. On three datasets for skeleton-based action recognition, our method achieves remarkable results with the same parameters as ST-GCN.

Related Works

Neural Networks with Graph. Graph neural network has attracted a lot of attention due to the effective representation for graph structure data such as social networks and the human skeleton. Existing GCN-models are mainly divided into two flows: The spectral-based method and the spatial-based method. The spectral-based GCNs (Defferrard, Bresson, and Vandergheynst 2016; Kipf and Welling 2016) transform graph signals to graph spectral domains based on eigen-decomposition, and perform convolution in the graph spectral domain, which is limited in terms of computational efficiency due to the heavy computation cost brought by the eigen-decomposition. In contrast to the spectral-based GCN, the spatial-based GCN (Simonyan and Zisserman 2014; Niepert, Ahmed, and Kutzkov 2016) is a generalization of the convolutional network from Euclidean space to non-Euclidean space, which directly implements convolution operation on the graph nodes and their neighbors in the spatial domain. This work follows the spatial-based method.

Skeleton-Based Action Recognition. As an alternative data source for RGB video, skeleton data has been widely used for action recognition due to the robustness against variations of backgrounds. The methods for skeleton-based human action recognition are mainly divided into two flows: the hand-crafted-based and the deep-learning-based. The hand-crafted-based methods usually focus on designing handcrafted features based on the physical intuitions such as joints angles (Ofli et al. 2014), distances (Xia, Chen, and Aggarwal 2012) and kinematic features (Zanfir, Leordeanu,

and Sminchisescu 2013). However, manually designing features requires a lot of energy, and it is difficult to consider all factors related to the action. With the development of deep learning, data-driven approaches which automatically learn the action patterns have attracted much attention. Some RNN-based models model the skeleton data as time sequences of coordinate vectors and capture temporal dynamics between consecutive frames, such as bi-RNNs (Du et al. 2015), deep LSTMs (Liu et al. 2016; Shahroudy et al. 2016), and attention-based model (Song et al. 2017). CNN-based models which try to embed the motion patterns into a skeleton image also achieve remarkable results (Ke et al. 2017; Soo Kim and Reiter 2017).

Recently, the graph-based methods draw much attention due to their superiority to model the relations between body joints. Yan, Xiong, and Lin proposed a spatial temporal graph convolutional network (ST-GCN) to directly model the skeleton data as graph structure (Yan, Xiong, and Lin 2018). To capture the relations between distant joints, some data-dependent methods were proposed. Shi et al.; Li et al.; Zhang, Xu, and Tao introduced additional mechanism to adaptively learn the relation between different joints (Shi et al. 2019c; Li et al. 2019a; Zhang, Xu, and Tao 2020). On the other hand, some approaches extracted multi-scale structural features via higher-order polynomials of the skeleton adjacency matrix. Li et al.; Peng et al.; Liu et al. introduced multiple-hop modules to break the limitation of representational capacity caused by one-order approximation (Li et al. 2019a; Peng et al. 2020; Liu et al. 2020). Different from their methods, we utilize a series of sub-graph convolutions cascaded by residual connections to capture both short-range joints dependencies and distant joints relations. For multi-scale temporal modeling, Liu et al. enhanced the regular temporal modeling with multi-scale learning and deployed parallel 3×1 kernel sizes with different dilation rates to enlarge the temporal receptive fields (Liu et al. 2020), while our model enriches the temporal receptive field within a single block and aggregates both short- and long-range temporal information via a hierarchical architecture.

Methodology

In this section, we first briefly review ST-GCN (Yan, Xiong, and Lin 2018) to make the paper self-contained. Then, we detail our MST-GCN for skeleton-based action recognition.

Basic Graph Convolution

A human skeleton graph is defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_{ti}|t = 1, \cdots, T, i = 1, \cdots, V\}$ includes all the joints in a skeleton sequence and $\mathcal{E} = \{\mathcal{E}_\mathcal{S}, \mathcal{E}_\mathcal{T}\}$ is the set of intra-skeleton and inter-frames edges. T denotes the number of frames and V denotes the number of nodes in a human skeleton.

In the spatial domain, a sampling function is defined as $\mathcal{B}_\mathcal{S}(v_{ti}) = \{v_{tj}|d(v_{tj}, v_{ti}) \leq D\}$ to determine the region range of the spatial graph convolution. Here $d(v_{tj}, v_{ti})$ denotes the shortest path from v_{tj} to v_{ti}, the parameter D controls the spatial range to be included in the neighbor graph. The direct natural connections in human skeleton

can be represented by the spatial adjacency matrix $A_\mathcal{S} \in R^{V \times V}$, where $A_{\mathcal{S}_{ij}} = 1$ if the joint i and j are directly connected and 0 otherwise. To capture refined location information, a spatial configuration labeling function is proposed in ST-GCN (Yan, Xiong, and Lin 2018). It divides the neighbor set into three subsets: (1) the root node itself; (2) the centripetal subset, which contains the neighboring nodes that are closer to the center of gravity than the root node; (3) the centrifugal subset, which contains the neighboring nodes that are farther from the center of gravity than the root node. Thus $A_\mathcal{S}$ is accordingly parted to be $A_\mathcal{S}^{root}$, $A_\mathcal{S}^{centripetal}$ and $A_\mathcal{S}^{centrifugal}$. We have 3 different subsets $\mathcal{P} = \{root, centripetal, centrifugal\}$.

For the temporal dimension, the sampling function is defined as $\mathcal{B}_\mathcal{T}(v_{ti}) = \{v_{qi}| |q - t| \leq \lfloor \Gamma/2 \rfloor\}$, where parameter Γ controls the temporal scales to aggregate. The temporal adjacency matrix is defined as $A_\mathcal{T} \in R^{T \times T}$ to represent the trajectories of the joints between consecutive frames. With a simple labeling function based on the temporal sequence, the temporal adjacency matrix can be partitioned into Γ parts: $A_\mathcal{T}^{-\lfloor \Gamma/2 \rfloor}, \cdots, A_\mathcal{T}^0, \cdots, A_\mathcal{T}^{\lfloor \Gamma/2 \rfloor}$.

With the defined graph, refined sampling and labeling function, the graph convolution operations on node v_{ti} are formulated as:

$$Y_\mathcal{S}(v_{ti}) = \sum_{v_{tj} \in \mathcal{B}_\mathcal{S}(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} X(v_{tj}) W_\mathcal{S}(\mathcal{L}_\mathcal{S}(v_{tj})), \quad (1)$$

$$Y_\mathcal{T}(v_{ti}) = \sum_{v_{tj} \in \mathcal{B}_\mathcal{T}(v_{ti})} \frac{1}{Z_{ti}(v_{qi})} X(v_{qi}) W_\mathcal{T}(\mathcal{L}_\mathcal{T}(v_{qi})), \quad (2)$$

where, $X(\cdot)$ is the feature of a node. $W_\mathcal{S}(\cdot)$ and $W_\mathcal{T}(\cdot)$ are weight functions that allocate a weight indexed by label $\mathcal{L}_\mathcal{S}(v_{tj})$ and $\mathcal{L}_\mathcal{T}(v_{qi})$ from a set of weights. $Z_{ti}(\cdot)$ is the number of the corresponding subset, which normalizes feature representations. $Y_\mathcal{S}(v_{ti})$ and $Y_\mathcal{T}(v_{ti})$ denote the outputs of spatial graph convolution and temporal graph convolution at node v_{ti}, respectively.

Implementation

The implementation of the spatial graph convolution is not straightforward. Specifically, we denote the feature map of the network is $X \in R^{C \times T \times V}$, where C denotes the number of channels. To implement the ST-GCN (Yan, Xiong, and Lin 2018), Eq. 1 is transformed into

$$Y_\mathcal{S} = \sum_{p \in \mathcal{P}} W_\mathcal{S}^{(p)}(X \hat{A}_\mathcal{S}^{(p)}) \odot M^{(p)}, \quad (3)$$

where, $\hat{A}_\mathcal{S}^{(p)} = D_\mathcal{S}^{(p)^{-\frac{1}{2}}} A_\mathcal{S}^{(p)} D_\mathcal{S}^{(p)^{\frac{1}{2}}}$ is a normalized adjacency matrix in spatial configuration of the p subset, and $D_\mathcal{S}^{(p)} = \sum_j A_{\mathcal{S}_{ij}}^{(p)} + \alpha$ is a degree matrix. α is set to 0.001 to avoid empty rows. $W_\mathcal{S}^{(p)} \in R^{C_{out} \times C_{in} \times 1 \times 1}$ is a trainable weight vector of the 1×1 convolution operation. $M^{(p)} \in R^{N \times N}$ is a mask map that indicates the importance of each joint. $\odot$ denotes the dot product. As mentioned in (Shi et al. 2019c,a), the dot multiplication operation can change only the importance of existing edges without adding new edges.

Therefore, we replace the dot product operation with an additional operation, the Eq. 3 is transformed into

$$Y_{\mathcal{S}} = \sum_{p \in \mathcal{P}} W_{\mathcal{S}}^{(p)} X (\hat{A}_{\mathcal{S}}^{(p)} + M^{(p)}), \tag{4}$$

Considering the consistency of the spatial temporal graph and the video in temporal dimension, we implement temporal graph convolution with classical convolution operation, and Eq. 2 can be represented as:

$$Y_{\mathcal{T}} = Conv2D[K_t \times 1](X), \tag{5}$$

where $Conv2D[K_t \times 1]$ is the classic 2D convolutional operation with the kernel size $K_t \times 1$.

In ST-GCN, D is set to 1 and K_t is set to 9, which results in the limited spatial temporal receptive field of the model and lack of flexibility. To solve these problems, we propose MS-GC and MT-GC modules.

Multi-Scale Spatial Graph Convolution Module

Recent studies show that capturing dependencies between both short-range nodes and long-range nodes is essential for skeleton based human action recognition (Li et al. 2019a; Shi et al. 2019c; Zhang, Xu, and Tao 2020; Liu et al. 2020). However, these methods either introduce additional modules to adaptively learn the relationship between different nodes or use higher-order polynomials of the skeleton adjacency matrix. They fail to well control the balance between performance and parameters. In this paper, we propose the MS-GC module, in which the spatial features and corresponding local graph convolution are split into a group of subsets. The main idea is inspired by Res2Net (Gao et al. 2019) which has a great impact on many fields of image processing. We transfer it to GCN-based action recognition with skeleton data. The architecture of the proposed MS-GC module is shown in Fig. 2. The subsets are formulated as a hierarchical residual architecture, so features can be hierarchically processed. The equivalent receptive field of the spatial dimension is enlarged and the model is capable of capturing the relations between those distant joints.

Specifically, given an input feature X whose shape is $[C, T, V]$, we split the feature into s fragments along the channel dimension, denoted by $\mathbf{x}_i$, where $i \in \{1, 2, \cdots, s\}$. The shape of each fragment thus becomes $[C/s, T, V]$. Each fragment $\mathbf{x}_i$ has a corresponding spatial graph convolution which is implemented using Eq. 4, denoted by $\mathbf{G}_i$. Each sub-spatial graph convolution has $1/s$ number of channels compared with the original ones and accordingly has only $1/s^2$ parameters as the original ones. Moreover, the residual connection is placed between two adjacent fragments, which enrich the diversity of receptive fields to capture the dependencies between both local and non-local joints. Formally,

$$\mathbf{y}_i = \begin{cases} \mathbf{G}_i(\mathbf{x}_i) & i = 1 \\ \mathbf{G}_i(\mathbf{x}_i + \mathbf{y}_{i-1}) & i > 1 \end{cases} \tag{6}$$

where $\mathbf{y}_i \in \mathbf{R}^{C/s \times T \times V}$ is the output of i_{th} sub-spatial graph convolution.

In this module, these s fragments have different receptive fields. For instance, with D set to 1, $\mathbf{G}_1$ can aggregate information from 1-hop neighbors, while $\mathbf{G}_2$ could potentially
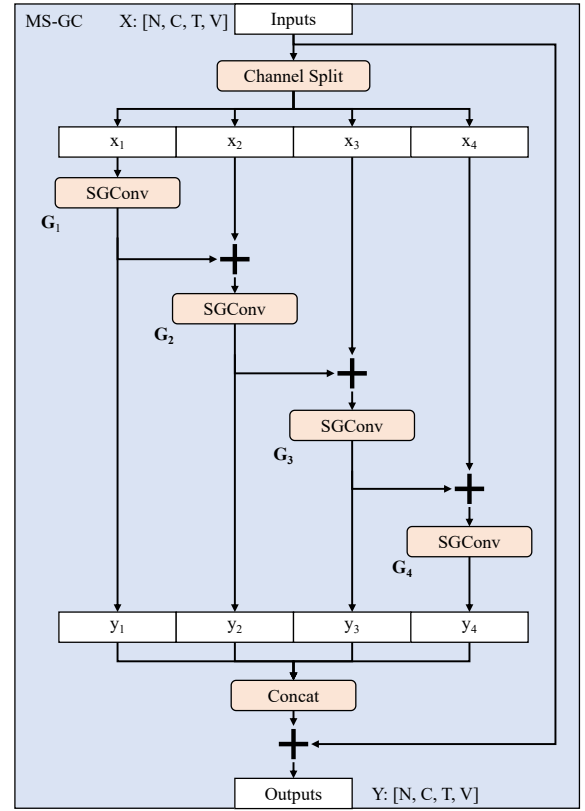
Figure 2: Illustration of multi-scale spatial graph convolution (MS-GC) module. SGConv denotes the spatial graph convolution, N is the batch size.

receive feature information from 2-hop neighbors by aggregating the information from $\mathbf{y}_1$. Thus the equivalent receptive field of the last fragment $\mathbf{y}_s$ has been enlarged several times. Finally, all the fragments are concatenated and an additional residual connection for the entire module is adopted to help the model converge. The outputs of MS-GC module can be computed as:

$$Y = \sigma([\mathbf{y}_1; \cdots; \mathbf{y}_s] + X) \tag{7}$$

where σ is the activation function. The obtained output feature Y involves capturing spatial feature representations between nodes at different distances which is superior to the local spatial representations obtained by using a single local graph convolution in typical approaches.

The MS-GC module is efficient: it can control the balance between complexity and the multi-scale representation ability of the model by adjusting s. Without introducing additional parameters and time-consuming operations, the module is able to capture the dependencies between both short- and long-range joints.

Multi-Scale Temporal Graph Convolution Module

Temporal modeling is essential to action recognition as well, while long-range temporal modeling is largely ignored in previous works. Long-range temporal dependence can not

only reduce the ambiguity between actions through temporal information, but also provide global information to help the model better learn spatial temporal features. Many existing works perform temporal modeling using temporal convolutions with a fixed kernel size throughout the architecture (Yan, Xiong, and Lin 2018; Li et al. 2019a; Shi et al. 2019c; Peng et al. 2020). The long-range temporal relationship is indirectly modeled by repeated stacking local temporal graph convolutions in deep networks. However, after a large number of local convolution operations, the useful features from distant frames have already been weakened and cannot be captured well.

To address this problem described above, we naturally extend the MS-GC module to the temporal domain. The proposed MT-GC module has a similar structure to MS-GC, but does not introduce additional residual connections. It replaces the local temporal graph convolution with a set of sub-temporal graph convolutions which are constructed as hierarchical residual-like connections. The temporal graph convolution in each subset is the same, but has different inputs. Similar to the MS-GC module, when the spatial temporal features go through the MT-GC module, a series of cascaded temporal graph convolution operations are applied to corresponding fragments to enlarge the temporal receptive field. Finally, a simple concatenation strategy is adopted to fuse the fragments. Therefore, the final output has the multi-scale temporal representation while both the short-range and long-range temporal relationships are well captured.

Multi-Scale Spatial Temporal Graph Convolution

To have a fair comparison with the state-of-the-art methods, we follow the same method as ST-GCN (Yan, Xiong, and Lin 2018) to construct our multi-scale spatial temporal graph convolution network. The ST-GCN backbone is shown as Fig. 3(a), which consists of 1 batch normalization layer to normalize the data and 10 ST-GC blocks to extract spatial and temporal features. Each ST-GC block, illustrated as Fig. 3(b), contains a spatial graph convolution and a followed temporal graph convolution to extract spatial and temporal features alternately. A global average pooling operation is adopted to aggregate the spatial temporal information and the final prediction results are output through the full connected layer together with a softmax layer.

Our work focuses on extracting discriminative spatial and temporal features, and there are two modes of combining the MS-GC module with the MT-GC module. In the first method, we replace the spatial graph convolution unit with our MS-GC module, and the temporal graph convolution is replaced with our MT-GC module as well. In the second method, we construct a spatial temporal residual graph convolution (STR-GC) module by concatenating the sub- spatial and temporal graph convolution within a single block, which is illustrated as Fig. 3(c). For convenience, we denote the sub-temporal graph convolution as $\mathbf{T}_i$. In this module, the spatial and temporal features will be updated alternately in each subset, and the corresponding spatial and temporal receptive field are enlarged due to the combined explosion
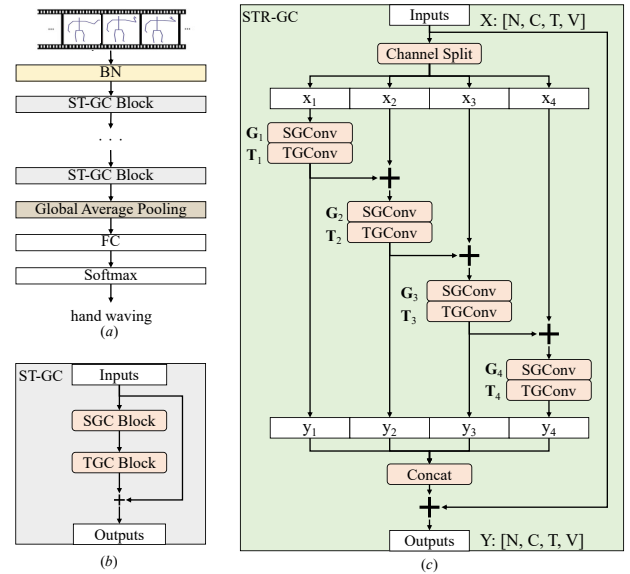

Figure 3: Architecture overview. (a) The full architecture of the ST-GCN backbone, BN is a batch normalization layer, FC is a full connected layer. (b) Illustration of ST-GC block which is used to build the whole network. (c) Illustration of proposed spatial temporal residual graph convolution (STR-GC) module, TGConv is the temporal graph convolution operation.

effect. Formally,

$$\mathbf{y}_i = \begin{cases} \mathbf{T}_i(\mathbf{G}_i(\mathbf{x}_i)) & i = 1 \\ \mathbf{T}_i(\mathbf{G}_i(\mathbf{x}_i + \mathbf{y}_{i-1})) & i > 1 \end{cases} \quad (8)$$

The first module can maintain the continuity of the model in extracting multi-scale temporal feature and spatial feature, while the latter is lighter than the first one, and it is more convenient to expand the spatial temporal joint learning that we leave to future work. The comparison of these two modules will be involved in the ablation study.

Experiments

Datasets

NTU RGB+D: NTU RGB+D (Shahroudy et al. 2016) dataset contains 60 different human action classes and it is the most widely used dataset for evaluating skeleton-based action recognition models. It consists of 56,880 action samples in total which are performed by 40 distinct subjects. The 3D skeleton data is collected by Microsoft Kinect v2 from three cameras simultaneously with different horizontal angles: -45°, 0°, 45°. The human pose in each frame is represented by 25 joints. There are two evaluation protocols for this dataset: Cross-Subject (X-sub) and Cross-View (X-view). Under the former protocols, half of the 40 subjects consists of the training set and the other for testing. For the latter, samples captured by camera 2 and 3 are used for training and the rest are for testing.

NTU-120 RGB+D: NTU-120 RGB+D (Liu et al. 2019) dataset is the currently largest 3D skeleton-based action recognition dataset captured under various environmental conditions. This dataset is an expansion of NTU RGB+D dataset in the number of performer and action categories. It contains 114,480 action samples from 120 action classes. Samples are captured by 106 subjects with three camera views. There are 32 setups that denote a specific location and background. Following the recommendation of the author, two benchmarks are used to evaluate the performance of models. Cross-Subject (X-sub) split the 106 subjects into training and testing group. Each group contains 53 subjects. Cross-setup (X-setup) divides samples with even setup IDs for training and odds setup IDs for testing.

Kinetics-Skeleton: Kinetics dataset (Kay et al. 2017) contains about 300,000 video clips in 400 classes collected from the Internet. The skeleton information is not provided by the original dataset but estimated by publicly available Open-Pose (Cao et al. 2017) toolbox. The captured skeleton information contains 18 body joints, along with their 2D coordinates and confidence score. There are 240,436 samples for training and 19794 samples for testing. Following the conventional evaluation method, Top-1 and Top-5 accuracies are reported.

Training details The model is built by stacking 10 building blocks, the first four blocks of the baseline model have 64 channels for output. In the 5_{th} and 8_{th} blocks, the number of channels is doubled while the temporal length is downsampled by a factor 2 according to the previous blocks. In the comparison experiment, the number of channels is adjusted to ensure a fair comparison under the same level of parameter complexity. We use SGD with Nesterov momentum (0.9), batch size 24, an initial learning rate 0.1 (can linearly scale up with batch size (Goyal et al. 2017)) to train the model for 110 epochs. The learning rate is decayed by 10 at 50_{th}, 70_{th} and 90_{th} epoch. For the frame which contains more than two persons, only the top-2 persons are selected based on the average energy. All skeleton sequences are padded to $T = 300$ frames by replaying the actions, but a sliding window of size 150 is used to randomly choose the crop input of sequence for Kinetics-Skeleton dataset. The same data pre-processing as in (Shi et al. 2019c) is adopted for NTU RGB+D and NTU-120 RGB+D datasets. All experiments in the ablation study used the above setting, including both our proposed method and the baseline model.

Ablation Study

In this subsection, we examine the effectiveness of the proposed MS-GC module, MT-GC module and their combination. The number of channels for each block will be adjusted to maintain the consistency of the parameters. The baseline model is the adaptive ST-GCN which is discussed in Sec. 3.2 implementation.

MS-GC module To verify the effectiveness of the proposed MS-GC module, we compare the MS-GCN with adaptive ST-GCN and perform the comparison experiments on NTU RGB+D with the benchmark of cross-subject. To build our MS-GCN, we replace the SGC block in 9 residual ST-GC blocks with our MS-GC module. The only difference between our MS-GC module and the SGC block is the set of s. When s set to 1, the MS-GC module is actually the SGC block, and when $s > 1$, the module gains the ability of multi-scale spatial representation. Here we first compare the number of the split subsets s with different settings. As shown in Table 1, with the increase of s, the MS-GCN achieves better performance with the similar parameters. Specifically, when s set to 4, our MS-GC module can improve the baseline at 1.0%, indicating that the importance of capturing the dependencies between distant joints and that our proposed MS-GC module can effectively enlarge the receptive field to capture such dependencies. To further explore the effectiveness of the MS-GC module, we also designed a lightweight MS-GCN to compare with other models. Each block of the lightweight MS-GCN has the same number of channels as ST-GCN, but the entire network has less parameters. Comparing the lightweight MS-GCN with the baseline model, our model achieves better performance. The improved performance demonstrates that our proposed MS-GC module is equipped with the ability to capture the relations between the local and non-local joints for better spatial representations.

	Setting	Params	Top 1
ST-GCN	$64c \times 1s$	3.1M	87.2%
MS-GCN	$34c \times 2s$	3.1M	87.8%
	$23c \times 3s$	3.0M	88.1%
	$17c \times 4s$	3.0M	**88.2%**
	$16c \times 4s$	**2.7M**	88.0%

Table 1: Comparisons between regular SGC and MS-GCN modules with difference of s on NTU RGB+D with X-sub benchmark. c denotes the output channels of the sub-SGCs in the second block. s denotes the number of subsets we divide the spatial features into.

MT-GC module To validate the efficacy of MT-GC module, we also compare our MT-GCN with the adaptive ST-GCN on NTU RGB+D cross-subject benchmark. We build our MT-GCN by replacing the TGC of ST-GCN with MT-GC module, and the only difference between our MT-GCN and the baseline is inserting our MT-GC module. Here we first evaluate the effectiveness and efficiency of different s. As shown in Table 2, when the number of split subsets is increased, we observe consistently better results with similar parameters, verifying MT-GC module is capable of capturing long-range temporal dynamics that are rarely considered in current work. Especially, when we split the inputs into 4 subsets along with the channel dimensions, we improve the accuracy by 1.2%. We also designed a lightweight MT-GCN which has the same channels setting for each block as the ST-GCN. Compared with the baseline model, the lightweight MT-GCN outperforms the ST-GCN with an absolute improvement of 0.7%, which demonstrates that our MT-GCN is efficient. All the above comparison experiments have demonstrated the effectiveness of our proposed mod-

	Setting	Params	Top 1
ST-GCN	$64c \times 1s$	3.1M	87.2%
MT-GCN	$40c \times 2s$	3.1M	87.7%
	$30c \times 3s$	3.1M	88.1%
	$24c \times 4s$	3.1M	**88.4%**
	$16c \times 4s$	**1.4M**	87.9%

Table 2: Comparisons between regular TGC and MT-GCN modules with difference of s on NTU RGB+D with X-sub benchmark. c is the output channels of the sub-TGCs in the second block. s is the number of subsets we divide the temporal features into.

ule. It has the ability to extract both short- and long-range temporal dependencies.

MST-GC module We conduct multi-scale spatial and temporal graph convolution and explore method to effectively combine the two proposed modules. According to previous ablation experiments, s is set to 4 for both MS-GC and MT-GC modules to achieve the best performance. The MST-GCN replaces both SGC and TGC in the ST-GC block with the MS-GC and MT-GC modules, respectively. While the STR-GCN replaces ST-GC blocks with the STR-GC module. As shown in Table 3, the MS-GC and MT-GC modules are complementary and their combination can promote each other to achieve better performance. STR-GCN is a model with potential, it achieves good performance with fewer parameters under the same settings. MST-GCN improves the adaptive ST-GCN at 1.8% with the similar parameters and 0.9% with only one third parameters, which validates the complementarity of these two module and also demonstrates that our MST-GCN can capture multi-scale spatial and temporal dependencies, leading to better motion patterns learning.

Spatial model	Temporal model	Setting	Params	Top 1
Regular SGC	Regular TGC	$64c \times 1s$	3.1M	87.2%
MS-GC	Regular TGC	$17c \times 4s$	3.1M	88.2%
Regular SGC	MT-GC	$24c \times 4s$	3.1M	88.4%
MS-GC	MT-GC	$16c \times 4s$	**0.9M**	88.1%
MS-GC	MT-GC	$30c \times 4s$	3.0M	**89.0%**
STR-GC		$30c \times 4s$	2.8M	88.5%

Table 3: The effectiveness and efficiency of MST-GCN. The accuracy is on NTU RGB+D cross-subject task. s is the number of subsets we divide both the spatial temporal features into.

Feature visualization. To validate how the features of each joint affect the final performance, we visualize the output feature maps of the last MST-GC block in Fig. 4, where the circle around each joint indicates magnitude of feature responses of this joint. As shown in Fig. 4, our model is capable of focusing on the parts that are most relevant to the action. Specifically, both the left arm and left foot among ac-
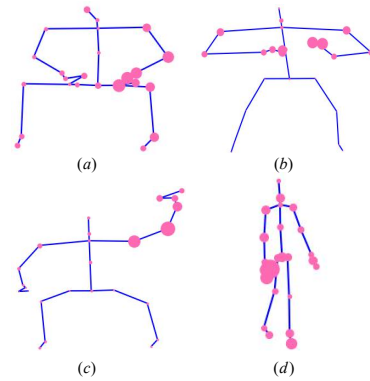

Figure 4: Feature responses in the last layer of MST-GCN backbone. The areas of circles indicate the response magnitudes. (a) "wearing a shoe" (b) "clapping" (c) "hand waving" (d) "walking"

tion "wearing a shoe" is well focused. Among action "clapping" and "hand waving", large responses are distributed on the arm parts, and there is almost no response from other parts, which could improve the robustness of the model and reduce the interference of noise nodes. The model could also capture the coordination of the whole body when walking. On the other hand, our MST-GCN could capture useful long-range dependencies to recognize the actions.

Comparison with State of the Arts

Considering the complementary information between different data, many state-of-the-art methods utilize multi-stream fusion strategies. To conduct a fair comparison, we adopt the same multi-stream fusion strategy as Shi et al. (2019a,b); Cheng et al. (2020), which utilizes the joint, bone, joint motion and bone motion streams. The "joint stream" uses the original skeleton coordinates as input, and the "bone stream" uses the the differential of spatial coordinates (the second order information of skeleton joints) as input. The "joint motion stream" and "bone motion stream" uses the temporal differential between adjacent frames of corresponding data as input. A simple score-level fusion strategy is adopted to obtain the fused score.

The results of 1-stream which only uses the joint or bone stream and 2-stream which utilizes the both joint and bone stream and 4-stream which uses all 4 streams are reported in our experiments. To verify the generality of our approach, we compare the MST-GCN with state-of-the-art methods on three different datasets: NTU RGB+D dataset, NTU-120 RGB+D dataset and Kinetics-Skeleton dataset, the corresponding results are shown in Table 4, Table 5, Table 6. On NTU RGB+D dataset, MST-GCN achieves competitive performance on cross-view and cross-subject benchmarks. For NTU-120 RGB+D dataset, MST-GCN achieves the-state-of-the-art results on both cross-subject and cross-setup benchmarks. For Kinetics-skeleton, MST-GCN achieves the-state-of-the-art result on Top-1 and competitive result on Top-5. The competitive and the state-of-the-art results verify the superiority of our model.

Methods	X-view (%)	X-sub (%)
HBRNN (Du et al. 2015)	64.0	59.1
P-LSTM (Shahroudy et al. 2016)	67.3	60.7
TCN (Soo Kim and Reiter 2017)	83.1	74.3
VA-LSTM (Zhang et al. 2017)	87.7	79.2
ST-GCN (Yan, Xiong, and Lin 2018)	88.3	81.5
AS-GCN (Li et al. 2019a)	94.2	94.2
2s AGC-LSTM (Si et al. 2019)	95.0	89.2
2s AGCN (Shi et al. 2019c)	95.1	88.5
2s NAS-GCN (Peng et al. 2020)	95.7	89.4
4s DGNN (Shi et al. 2019a)	96.1	89.9
4s MS-AAGCN (Shi et al. 2019b)	96.2	90.0
2s MS-G3D (Liu et al. 2020)	96.2	**91.5**
4s Shift-GCN (Cheng et al. 2020)	96.5	90.7
Js MST-GCN (ours)	95.1	89.0
Bs MST-GCN (ours)	95.2	89.5
2s MST-GCN (ours)	96.4	91.1
4s MST-GCN (ours)	**96.6**	**91.5**

Table 4: Comparisons of the Top-1 accuracy with the state-of-the-art methods on the NTU RGB+D dataset.

Methods	X-sub (%)	X-setup (%)
ST-LSTM (Liu et al. 2016)	55.7	57.9
MTCNN + RotClips (Ke et al. 2018)	62.2	61.8
SkeMotion (Liu et al. 2017)	67.7	66.9
TSRJI (Caetano et al. 2019)	67.9	62.8
4s Shift-GCN (Cheng et al. 2020)	85.9	87.6
2s MS-G3D (Liu et al. 2020)	86.9	88.4
Js MST-GCN (ours)	82.8	84.5
Bs MST-GCN (ours)	84.8	86.3
2s MST-GCN (ours)	87.0	88.3
4s MST-GCN (ours)	**87.5**	**88.8**

Table 5: Comparisons of the Top-1 accuracy with the state-of-the-art methods on the NTU-120 RGB+D dataset.

Methods	Top-1 (%)	Top-5 (%)
ST-GCN (Yan, Xiong, and Lin 2018)	31.6	53.7
AS-GCN (Li et al. 2019a)	34.8	56.3
2s AGCN (Shi et al. 2019c)	36.1	58.7
4s DGNN (Shi et al. 2019a)	36.9	56.9
2s NAS-GCN (Peng et al. 2020)	37.1	60.1
2s MS-G3D (Liu et al. 2020)	38.0	**60.9**
Js MST-GCN (ours)	36.0	58.5
Bs MST-GCN (ours)	32.3	58.2
2s MST-GCN (ours)	37.8	60.3
4s MST-GCN (ours)	**38.1**	60.8

Table 6: Comparisons of the Top-1 and Top-5 accuracy with the state-of-the-art methods on the Kinetics dataset.

Discussion

As discussed in ST-GCN (Yan, Xiong, and Lin 2018), the skeletons can provide complementary information to RGB-based and optical flow-based models. We also explore us-

	RGB TSN	ST-GCN	MST-GCN	Top 1 (%)
Single Model	√			70.4
		√		31.6
			√	36.0
Ensemble Model	√	√		71.2
	√		√	72.1

Table 7: Combining RGB model with different skeleton models. The accuracy is on Kinetics validation set.

ing MST-GCN to capture motion information in two-stream style action recognition. The RGB model is a standard temporal segment networks (TSN) (Wang et al. 2016) with pre-trained weights[1]. For fair comparison, both ST-GCN and MST-GCN only utilize the joint stream. As shown in Table 7, there is a big gap between the RGB-based models and the skeleton-based models, but the skeleton-based models can provide useful motion information to boost the performance of the RGB-based models. Adding MST-GCN to the RGB-based model leads to 1.7% increase, much better than ST-GCN (0.8%). These results clearly show that the improvement obtained in skeleton-based models could further improve the performance of the ensemble model.

Conclusion

In this work, we propose a MS-GC module and a MT-GC module to capture both short- and long-range dependencies in spatial and temporal domains, respectively. The MS-GC and MT-GC modules enhance the multi-scale representation ability of the model by decomposing a local convolution into several sub-convolutions and combining them through residual connections. Both of them are simple but efficient, they are also complementary to each other. By coupling these methods, we propose a multi-scale spatial and temporal graph convolution network (MST-GCN), which is a powerful network to extract multi-scale spatial and temporal features. On three large-scale challenging datasets, the proposed MST-GCN achieves competitive or the-state-of-the-art results. Considering the relation between the spatial and temporal, jointly extracting spatial-temporal features is very helpful to explore the motion pattern in the spatial-temporal domain. In the future, we will try the spatial and temporal joint learning to promote the performance of human action recognition. We will further explore the promotion of the skeleton-based methods to RGB-based methods.

Ackonwledgements

---

[1]https://github.com/open-mmlab/mmaction2

References

Aggarwal, J. K.; and Ryoo, M. S. 2011. Human Activity Analysis: A Review. *ACM Computing Surveys (CSUR)* 43(3): 1–43.

Caetano, C.; Brémond, F.; Schwartz, W. R.; and and. 2019. Skeleton Image Representation for 3D Action Recognition Based on Tree Structure and Reference Joints. In *Proceedings of the SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 16–23.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7291–7299.

Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; and Lu, H. 2020. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 183–192.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Advances in Neural Information Processing Systems (NIPS)*, 3844–3852.

Du, Y.; Wang, W.; Wang, L.; and and. 2015. Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1110–1118.

Gao, S.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; and Torr, P. H. 2019. Res2net: A Mew Multi-Scale Backbone Architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* .

Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, Large Minibatch SGD: Training Imagenet in 1 Hour. *arXiv preprint arXiv:1706.02677* .

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950* .

Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; and Boussaid, F. 2017. A New Representation of Skeleton Sequences for 3D Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3288–3297.

Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; and Boussaid, F. 2018. Learning Clip Representations for Skeleton-Based 3D Action Recognition. *IEEE Transactions on Image Processing (TIP)* 27(6): 2842–2855.

Kipf, T. N.; and Welling, M. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907* .

Li, C.; Zhong, Q.; Xie, D.; and Pu, S. 2017. Skeleton-Based Action Recognition with Convolutional Neural Networks. In *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 597–600.

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2019a. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3595–3603.

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2019b. Symbiotic Graph Neural Networks for 3D Skeleton-Based Human Action Recognition and Motion Prediction. *arXiv preprint arXiv:1910.02212* .

Li, Y.; Ji, B.; Shi, X.; Zhang, J.; Kang, B.; and Wang, L. 2020. TEA: Temporal Excitation and Aggregation for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 909–918.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2019. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* .

Liu, J.; Shahroudy, A.; Xu, D.; and Wang, G. 2016. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 816–833.

Liu, J.; Wang, G.; Hu, P.; Duan, L.-Y.; and Kot, A. C. 2017. Global Context-Aware Attention LSTM Networks for 3D Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1647–1656.

Liu, M.; Liu, H.; and Chen, C. 2017. Enhanced Skeleton Visualization for View Invariant Human Action Recognition. *Pattern Recognition (PR)* 68: 346–362.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 143–152.

Niepert, M.; Ahmed, M.; and Kutzkov, K. 2016. Learning Convolutional Neural Networks for Graphs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014–2023.

Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; and Bajcsy, R. 2014. Sequence of The Most Informative Joints (smij): A New Representation for Human Skeletal Action Recognition. *Journal of Visual Communication and Image Representation* 25(1): 24–38.

Peng, W.; Hong, X.; Chen, H.; and Zhao, G. 2020. Learning Graph Convolutional Network for Skeleton-Based Human Action Recognition by Neural Searching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2669–2676.

Poppe, R. 2010. A Survey on Vision-Based Human Action Recognition. *Image and Vision Computing* 28(6): 976–990.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB + D: A Large Scale Dataset for 3D Human Activity Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1010–1019.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019a. Skeleton-Based Action Recognition with Directed Graph Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7912–7921.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019b. Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks. *arXiv preprint arXiv:1912.06971* .

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019c. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12026–12035.

Si, C.; Chen, W.; Wang, W.; Wang, L.; and Tan, T. 2019. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1227–1236.

Simonyan, K.; and Zisserman, A. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems (NIPS)*, 568–576.

Song, S.; Lan, C.; Xing, J.; Zeng, W.; and Liu, J. 2017. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4263–4270.

Soo Kim, T.; and Reiter, A. 2017. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 20–28.

Sudha, M.; Sriraghav, K.; Jacob, S. G.; Manisha, S.; et al. 2017. Approaches and Applications of Virtual Reality and Gesture Recognition: A Review. *International Journal of Ambient Computing and Intelligence (IJACI)* 8(4): 1–18.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4489–4497.

Tu, J.; Liu, M.; and Liu, H. 2018. Skeleton-Based Human Action Recognition Using Spatial Temporal 3D Convolutional Neural Networks. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Vemulapalli, R.; Arrate, F.; and Chellappa, R. 2014. 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 (CVPR)*, 588–595.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, 20–36. Springer.

Wang, P.; Li, W.; Ogunbona, P.; Wan, J.; and Escalera, S. 2018. RGB-D-Based Human Motion Recognition With Deep Learning: A Survey. *Computer Vision and Image Understanding* 171: 118–139.

Weinland, D.; Ronfard, R.; and Boyer, E. 2011. A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition. *Computer Cision and Image Understanding* 115(2): 224–241.

Xia, L.; Chen, C.-C.; and Aggarwal, J. K. 2012. View Invariant Human Action Recognition Using Histograms of 3D Joints. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 20–27.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 7444–7452.

Zanfir, M.; Leordeanu, M.; and Sminchisescu, C. 2013. The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2752–2759.

Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; and Wang, H. 2016. Real-Time Action Recognition with Enhanced Motion Vector CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2718–2726.

Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; and Zheng, N. 2017. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition From Skeleton Data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2117–2126.

Zhang, X.; Xu, C.; and Tao, D. 2020. Context Aware Graph Convolution for Skeleton-Based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 14333–14342.