

# Spatio-Temporal Graph Convolution for Skeleton Based Action Recognition

Chaolong Li,\* Zhen Cui,\* Wenming Zheng, Chunyan Xu, Jian Yang

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

School of Biological Science & Medical Engineering, Southeast University, Nanjing, China

{zhen.cui, cyx, csjyang}@njust.edu.cn; {lichalong, wenming.zheng}@seu.edu.cn

## Abstract

Variations of human body skeletons may be considered as dynamic graphs, which are generic data representation for numerous real-world applications. In this paper, we propose a spatio-temporal graph convolution (STGC) approach for assembling the successes of local convolutional filtering and sequence learning ability of autoregressive moving average. To encode dynamic graphs, the constructed multi-scale local graph convolution filters, consisting of matrices of local receptive fields and signal mappings, are recursively performed on structured graph data of temporal and spatial domain. The proposed model is generic and principled as it can be generalized into other dynamic models. We theoretically prove the stability of STGC and provide an upper-bound of the signal transformation to be learnt. Further, the proposed recursive model can be stacked into a multi-layer architecture. To evaluate our model, we conduct extensive experiments on four benchmark skeleton-based action datasets, including the large-scale challenging NTU RGB+D. The experimental results demonstrate the effectiveness of our proposed model and the improvement over the state-of-the-art.

## 1 Introduction

Human action recognition is one of the most active research topics due to its wide applications to video surveillance, robot vision, human computer interaction, etc. Since human body itself can be viewed as an articulated system of rigid bones connected by hinged joints, the actions of human body are essentially embodied in skeletal motions in the 3D space (Ye et al. 2013). Thereby, various skeleton based action recognition methods (Liu et al. 2017; Shahroudy et al. 2016; Zhang et al. 2017) are springing up in recent years, accompanying with the progress of more accessible deep sensors.

Different from grid-shaped structures of images/videos, human skeleton, consisting of a series of joints and bones, has an irregular geometric structure. Human action may be regarded as a consecutive dynamic sequence of such irregular structures. Considering the motion complexity of the entire body skeleton, the sophisticated strategy is to separately model the trajectory of each joint or clique of joints

(body part). The statistic model (Xia, Chen, and Aggarwal 2012), shape model (Chaudhry et al. 2013), and geometric model (Vemulapalli, Arrate, and Chellappa 2014) are often used to characterize the motion trajectories. With the thriving of representation learning, the learning based methods, such as Hidden Markov Models (HMMs) (Xia, Chen, and Aggarwal 2012) and recursive models (long-short term memory, LSTM) (Liu et al. 2016; 2017; Song et al. 2017), become predominant in the dynamic representation of skeletal joints, because they dedicate more promising results on those public skeleton datasets.

Recently graph as a tool is used to represent skeletons (Wang et al. 2016a), although graph is popular to model various structured objects. Nevertheless, the graph based method (Wang et al. 2016a) still takes the conventional technique line of graph kernel matching. In fact, in the research fields of graph classification/matching, the graph representation (*e.g.*, the statistic on graphlet (Przulj 2007)) and graph metric (*e.g.*, graph kernel (Vishwanathan et al. 2010)) have been historically well-studied. With recent successes of deep learning on various problems, deep representation of graphs has aroused more attention (Yanardag and Vishwanathan 2015; Seo et al. 2016; Li et al. 2017). But the most crucial problem is the definition/identification of homogeneous graphs because the same responses should be produced from those homogeneous graphs. To this end, Niepert *et al.* (Niepert, Ahmed, and Kutzkov 2016) used a greedy strategy of sorting those nodes within a local neighbor region and then performed convolution-like filtering on the sorted nodes. Instead of this explicit spatial definition way, Defferrard *et al.* (Defferrard, Bresson, and Vandergheynst 2016) introduced a deep graph method based on spectral filtering, inspired by the recent signal processing theory on graph (Shuman et al. 2013). Li *et al.* (Li et al. 2017) introduced attention mechanism into graph convolution model. But those methods intrinsically belong to still-graph deep learning.

For dynamic graphs, some variants of recurrent neural network (RNN) are developed recently based on the traverse way of spatio-temporal graph. Jain *et al.* (Jain et al. 2016) proposed a structural-RNN by casting spatio-temporal graph as a RNN mixture for the task of action prediction. Li *et al.* (Li et al. 2015) proposed gated graph sequential neural network for the basic logical reasoning task. Seo *et*

\*Chaolong Li and Zhen Cui have equal contributions.

al. (Seo et al. 2016) fed the spatial filtered graph signals into LSTM for image generation. However, these methods do not yet absorb the essential successes of convolutional neural networks (CNN), which have changed AI landscape with breakthrough results on numerous applications.

In this paper, we propose a spatio-temporal graph convolution (STGC) approach to represent dynamic skeletal graph sequences. To encode the graph structure data, we design multi-scale convolutional filters, each of which is composed of receptive field computation and signal mapping. The receptive field is computed from the polynomial of adjacency matrix, which makes sure the same responses for homogeneous graphs. We simultaneously perform local convolutional filtering on temporal motions and spatial structures. The temporal convolutional filtering recursively encodes motion variations while the spatial filtering extracts more robust feature of spatial structures. In theory, the frequency responses of multi-scale convolutional filtering are equivalent to an approximate graph Fourier transform following by one linear function of feature mapping. Taking the philosophy of multi-scale convolutional filtering, we develop a recursive graph convolution model inspired by autoregressive moving average (ARMA). We theoretically analyze the stability of the proposed model and provide a theoretical upper-bound. Finally, we stack the spatio-temporal graph convolution into a deep architecture. To verify our proposed method, we conduct extensive experiments on four benchmark skeleton-based action datasets including the largest NTU RGB+D dataset (Shahroudy et al. 2016). The experimental results demonstrate the effectiveness of our proposed model and a more promising direction for skeleton based action recognition.

In summary, our contributions are three folds:

- Propose a spatio-temporal graph convolution approach, which assembles the successes of local convolutional filtering and sequence learning ability of recursive learning. The generic model is further extended to a deep architecture.
- Theoretically prove the stability of the proposed model and provide a theoretical upper-bound.
- Achieve the state-of-the-art performances on the four benchmark datasets including the large-scale challenging dataset NTU RGB+D.

## 2 Graph Preliminary

Consider an undirected graph  $\mathcal{G} = (\mathcal{V}, \mathbf{A})$  of  $N$  nodes, where  $\mathcal{V} = \{v_i\}_{i=1}^N$  is the set of nodes,  $\mathbf{A}$  is a (weighted) adjacency matrix. The adjacency matrix  $\mathbf{A} \in \mathbb{R}^{N \times N}$  records the connections between nodes, where if  $v_i, v_j$  are not connected, then  $A_{ij} = 0$ , otherwise  $A_{ij} \neq 0$ .

The graph Laplacian matrix  $\mathbf{L}$  is defined as  $\mathbf{L} = \mathbf{D} - \mathbf{A}$ , where  $\mathbf{D} \in \mathbb{R}^{N \times N}$  is the diagonal degree matrix with  $D_{ii} = \sum_j A_{ij}$ . A popular option is the normalized version, i.e., each weight  $A_{ij}$  is multiplied by a factor  $\frac{1}{\sqrt{D_{ii}D_{jj}}}$ , formally,  $\mathbf{L}^{norm} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}$ , where  $\mathbf{I}$  is the identity matrix. Unless otherwise specified, we use the normalized Laplacian matrix below.

As a symmetric semi-positive definite (SPD) matrix, the graph Laplacian  $\mathbf{L}$  has a complete set of orthonormal eigenvectors  $\{\phi_1, \dots, \phi_N\}$  satisfying  $\mathbf{L}\phi_i = \lambda_i\phi_i$ , where  $\{\lambda_i\}$  are nonnegative real eigenvalues. We assume all eigenvalues are ordered as  $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N = \lambda_{max}$ . For the normalized Laplacian matrix, we have a bound of  $\lambda_{max} = 2$ . In matrix expression, the Laplacian matrix can be written as  $\mathbf{L} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^\top$ , where  $\mathbf{\Lambda} = \text{diag}([\lambda_1, \lambda_2, \dots, \lambda_N])$  and  $\mathbf{\Phi} = [\phi_1^\top; \phi_2^\top; \dots; \phi_N^\top]^\top$ .

According to graph theory (Shuman et al. 2013), the graph Fourier transform (GFT) of a signal  $\mathbf{x}$  in spatial domain can be defined as  $\hat{\mathbf{x}} = \mathbf{\Phi}^\top \mathbf{x}$ , where  $\hat{\mathbf{x}}$  is the produced graph frequency signal. The corresponding inverse GFT is  $\mathbf{x} = \mathbf{\Phi}\hat{\mathbf{x}}$ . A graph filtering  $\mathbf{H}$  is an operator that acts upon a graph signal  $\mathbf{x}$  by amplifying or attenuating its graph Fourier coefficients as  $\mathbf{H}\mathbf{x} = \sum_{n=1}^N H(\lambda_n)\hat{x}_n\phi_n$ . The graph frequency response  $H : [\lambda_{min}, \lambda_{max}] \rightarrow \mathbb{C}$  controls how much  $\mathbf{H}$  amplifies the signal spectra  $H(\lambda_n) = (\phi_n^\top \mathbf{H}\mathbf{x})/\hat{x}_n$ .

## 3 The Model

Human body skeleton is represented with a group of 3D spatial coordinates of joints. Hence, one can use a graph to depict spatial relation of skeletal joints. To keep the original coordinate information, we add attributes of nodes into the graph, i.e.,  $\mathcal{G} = (\mathcal{V}, \mathbf{A}, \mathbf{X})$ , where  $\mathcal{V} = \{v_i\}_{i=1}^N$  is the set of nodes w.r.t. skeletal joints,  $\mathbf{A}$  is a (weighted) adjacency matrix and  $\mathbf{X}$  is a matrix of graph signals/attributes. According to the body bones between joints, we define those connected edges, and simply assign them to 1. In addition, other strategies (e.g., Gaussian kernel) may be used to produce the adjacency matrix. The signal matrix  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d] \in \mathbb{R}^{N \times d}$  is supported on the node set  $\mathcal{V}$ , whose  $i$ -th component (or channel)  $\mathbf{x}_i$  represents a signal of all nodes. That means, the  $i$ -th node  $v_i \in \mathcal{V}$  is assigned with a signal vector of  $d$  dimensions. For skeletal data, we define the signal of each joint with its 3D spatial coordinates, i.e.,  $\mathbf{x}_i : \mathcal{V} \rightarrow (x_i, y_i, z_i)$ . Thus, for a dynamic graph sequence of length  $T$ , we may formulate it as a stream of graphs  $(\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_T)$ , where  $\mathcal{G}_k = (\mathcal{V}, \mathbf{A}_k, \mathbf{X}_k)$  denotes the skeleton at the  $k$ -th time slice.

**Multi-scale Graphical Convolutional Kernels** In the standard CNN running on images, the receptive field may be conveniently defined as a local square spatial region, due to grid-shaped structure. So convolutional filtering on regular structures is accessible. On the contrary, the construction of convolutional kernels on graphs is intractable because the same filtering responses are required for homogeneous graph structures. Inspired by the graph theory (Shuman et al. 2013), we resort to the adjacency matrix  $\mathbf{A}$ , which expresses the connections between nodes. As  $\mathbf{A}^k$  exactly records the  $k$ -path reachable nodes, so we can construct a  $k$ -neighbor receptive field by defining a  $k$ -order polynomial of  $\mathbf{A}$ , denoted as  $\psi_k(\mathbf{A})$ . Taking the simplest case, let  $\psi_k(\mathbf{A}) = \mathbf{A}^k$ , which actually describes the  $k$ -hop neighbor nodes. In practice, we may replace  $\mathbf{A}$  with the Laplacian matrix  $\mathbf{L}$  to avoid the scale effect of matrix norm during the recursive inference (see the following model). Thus, for receptive fields of

$K$  scales, we define multi-scale convolutional filtering as

$$\mathbf{Z} = \mathcal{G} * f = \sum_{k=0}^{K-1} \psi_k(\mathbf{L}) \mathbf{X} \mathbf{V}_k, \quad (1)$$

where  $\psi_k(\mathbf{L})$  expresses the receptive field of the  $k$ -th scale and  $\mathbf{V}_k \in \mathbb{R}^{d \times d'}$  is the corresponding signal transformation. The computation of  $\psi_k(\mathbf{L}) \mathbf{X}$  weightedly summarizes the information of all nodes within the  $k$ -scale receptive field, which thus is homogenous-invariant for graphs.

**Spatio-Temporal Graph Convolution** Inspired by the design philosophy of autoregressive moving average (ARMA) (Hannan and Deistler 2012), we construct the spatio-temporal graphical convolutional model as follows,

$$\mathbf{Y}_{t+1} = \sum_{k=0}^{K_1-1} \psi_k(\mathbf{L}) \mathbf{Y}_t \mathbf{W}_k + \mathbf{X}_t \mathbf{V}_0, \quad (2)$$

$$\mathbf{O}_{t+1} = \mathbf{Y}_{t+1} + \sum_{k=1}^{K_2-1} \psi_k(\mathbf{L}) \mathbf{X}_t \mathbf{V}_k, \quad (3)$$

where  $\psi_k(\cdot)$  is a receptive field function on the  $k$ -th scale,  $\{\mathbf{W}_k \in \mathbb{R}^{d' \times d'}, \mathbf{V}_k \in \mathbb{R}^{d \times d'}\}$  are the signal transformation matrices with regard to the  $k$ -scale,  $K_1$  and  $K_2$  are the number of kernels respectively in the temporal and spatial domain. In the above model,  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_{d'}]$ ,  $\mathbf{O} = [\mathbf{o}_1, \dots, \mathbf{o}_{d'}]$  can be respectively viewed as the hidden state and the output state. Along time slices, signals are recursively regressed with local convolutional kernels in Eqn. (2), thus motion variations can be sequentially encoded. The output signals in Eqn. (3) combine spatially convolutional graph signals as well as dynamic temporal signals. Moreover, each output signal  $\mathbf{o}_i$  is dependent on all input signals  $\{\mathbf{x}_1, \dots, \mathbf{x}_d\}$ . Specifically, when signals are independent on each other, and the spatio-temporal convolutional filters are separately operated on a channel of signals, then the dynamic graph convolution model can be written as

$$\mathbf{Y}_{t+1} = \sum_{k=0}^{K_1-1} \psi_k(\mathbf{L}) \mathbf{Y}_t \text{diag}(\mathbf{w}_k) + \mathbf{X}_t \text{diag}(\mathbf{v}_0), \quad (4)$$

$$\mathbf{O}_{t+1} = \mathbf{Y}_{t+1} + \sum_{k=1}^{K_2-1} \psi_k(\mathbf{L}) \mathbf{X}_t \text{diag}(\mathbf{v}_k), \quad (5)$$

where  $\mathbf{w}_k = [w_{k1}, \dots, w_{kd}]^\top$ ,  $\mathbf{v}_k = [v_{k1}, \dots, v_{kd}]^\top$  are mapping parameters, and  $w_{ki}, v_{ki}$  are associated to the  $k$ -th scale of the  $i$ -th signal. Note here the dimension of output is assumed to be the input dimension. It is easy to extend into  $d' \neq d$  in the case of signal independency.

**Frequency Domain Analysis** As the receptive field function  $\psi_k(\cdot)$  is a polynomial expression, we can derive  $\psi_k(\mathbf{L}) = \psi_k(\phi \text{diag}(\boldsymbol{\lambda}) \phi^\top) = \phi \psi_k(\text{diag}(\boldsymbol{\lambda})) \phi^\top$ . Thus the frequency response of convolution filtering in Eqn. (1) can be written as

$$\widehat{\mathbf{Z}} = \sum_{k=0}^{K-1} \text{diag}(\psi_k(\boldsymbol{\lambda})) \widehat{\mathbf{X}} \mathbf{V}_k. \quad (6)$$

In particular, if signals are independent as expressed in Eqn. (4) and (5), then we can obtain the frequency responses of graphical signals as  $\widehat{\mathbf{Z}} = \sum_{k=0}^{K-1} (\psi_k(\boldsymbol{\lambda}) \mathbf{v}_k^\top) \odot \widehat{\mathbf{X}}$ , each frequency response of which is formally  $\widehat{Z}_{ij} = \sum_{k=0}^{K-1} (v_{kj} \psi_k(\lambda_i)) \widehat{X}_{ij}$ . Therefore, when signals are independent, the multi-scale convolutional filtering on graphs may be regarded as a  $(K-1)$ -order polynomial approximation of graph Fourier transform, if let  $H(\lambda_i) = \sum_{k=0}^{K-1} v_{ki} \psi_k(\lambda_i)$ .

In the signal dependency case, we can decompose  $\mathbf{V}_k$  into a diagonal matrix multiplied by a general matrix, i.e.,  $\mathbf{V}_k = \text{diag}(\boldsymbol{\alpha}_k) \widetilde{\mathbf{V}}_k$ . Then the frequency response of convolution filtering in Eqn. (6) can be written as  $\widehat{\mathbf{Z}} = \sum_{k=0}^{K-1} ((\psi_k(\boldsymbol{\lambda}) \boldsymbol{\alpha}_k^\top) \odot \widehat{\mathbf{X}}) \widetilde{\mathbf{V}}_k$ . Therefore, in the general case of signal dependency, the calculation of frequency responses may be understood as two steps: (i) perform a polynomial approximation of GFT on each input signal, and (ii) transform multi-channel signals by the new mappings  $\{\widetilde{\mathbf{V}}_k\}$ .

**Stability Analysis** For any sequence of graph realizations  $\{\mathcal{G}_1, \mathcal{G}_2, \dots\}$ , we can prove the stability of the recursive model (Eqn. (2)~(3)), which is summarized in the following theory.

**Theorem 1.** *Suppose the Laplacian matrix  $\mathbf{L}$  has the eigenvalue decomposition  $\mathbf{L} = \boldsymbol{\Phi} \boldsymbol{\Lambda} \boldsymbol{\Phi}^\top$ , and  $\psi_k(\cdot)$  is a  $k$ -order polynomial function satisfying  $\|\psi_k(\mathbf{L})\|_2 \leq 1$ . For the signal mappings  $\{\mathbf{W}_0, \dots, \mathbf{W}_{K_1-1}\}$ , if their diagonal elements are non-negative, i.e.,  $W_{k,ii} \geq 0$ , and  $\sum_{k=0}^{K_1-1} \|\mathbf{W}_k\|_\infty < 1$ , the frequency response  $\text{vec}(\widehat{\mathbf{O}})$  in Eqn. (3) will converge to*

$$\lim_{t \rightarrow \infty} \text{vec}(\widehat{\mathbf{O}}_t) = \mathcal{T} \cdot \text{vec}(\widehat{\mathbf{X}}_{t-1}), \quad (7)$$

$$\mathcal{T} = (\mathbf{I} - \Gamma_0^{K_1}(\mathbf{W}, \boldsymbol{\Lambda}))^{-1} \Gamma_0^1(\mathbf{V}, \boldsymbol{\Lambda}) + \Gamma_1^{K_2}(\mathbf{V}, \boldsymbol{\Lambda}), \quad (8)$$

where  $\Gamma_a^b(\mathbf{W}, \boldsymbol{\Lambda}) = \sum_{k=a}^{b-1} (\mathbf{W}_k^\top \otimes \psi_k(\boldsymbol{\Lambda}))$  (similar for  $\Gamma_0^{K_2}$ ). Moreover, the transformation function  $\mathcal{T}$  has an upper-bound:

$$\|\mathcal{T}\|_2 < \frac{\|\mathbf{V}_0\|_\infty}{1 - \sum_{k=0}^{K_1-1} \|\mathbf{W}_k\|_\infty} + \sum_{k=1}^{K_2} \|\mathbf{V}_k\|_\infty. \quad (9)$$

*Proof.* As  $\psi_k(\cdot)$  is a polynomial function, we can transform the recursive model (Eqn. (2)~(3)) from spatial domain into frequency domain by using  $\widehat{\mathbf{Y}} = \boldsymbol{\Phi}^\top \mathbf{Y}$  and  $\widehat{\mathbf{X}} = \boldsymbol{\Phi}^\top \mathbf{X}$ ,

$$\widehat{\mathbf{Y}}_{t+1} = \sum_{k=0}^{K_1-1} \psi_k(\boldsymbol{\Lambda}) \widehat{\mathbf{Y}}_t \mathbf{W}_k + \widehat{\mathbf{X}}_t \mathbf{V}_0, \quad (10)$$

$$\widehat{\mathbf{O}}_{t+1} = \widehat{\mathbf{Y}}_{t+1} + \sum_{k=1}^{K_2-1} \psi_k(\boldsymbol{\Lambda}) \widehat{\mathbf{X}}_t \mathbf{V}_k. \quad (11)$$

By using the abbreviated notation  $\Gamma_a^b(\cdot, \cdot)$  and  $\psi_0(\cdot) = \mathbf{I}$ , we can derive a vectorized style of Eqn. (10) as

$$\begin{aligned} \text{vec}(\widehat{\mathbf{Y}}_{t+1}) &= \Gamma_0^{K_1}(\mathbf{W}, \boldsymbol{\Lambda}) \text{vec}(\widehat{\mathbf{Y}}_t) + \Gamma_0^1(\mathbf{V}, \boldsymbol{\Lambda}) \text{vec}(\widehat{\mathbf{X}}_t) \\ &= \sum_{\tau=0}^t ((\Gamma_0^{K_1}(\mathbf{W}, \boldsymbol{\Lambda}))^\tau) \Gamma_0^1(\mathbf{V}, \boldsymbol{\Lambda}) \text{vec}(\widehat{\mathbf{X}}_t) \\ &\quad + (\Gamma_0^{K_1}(\mathbf{W}, \boldsymbol{\Lambda}))^{t+1} \text{vec}(\widehat{\mathbf{Y}}_0). \end{aligned} \quad (12)$$

Next we need to prove  $\|\Gamma_0^{K_1}(\mathbf{W}, \mathbf{\Lambda})\|_2 < 1$ . The matrix  $\Gamma_0^{K_1}(\mathbf{W}, \mathbf{\Lambda})$  is a block matrix of  $d' \times d'$  blocks, each of which is a diagonal matrix. Let  $\mathbf{A}^{(ij)} = \sum_{k=0}^{K_1-1} W_{k,ij} \psi_k(\mathbf{\Lambda})$  denote the  $(i, j)$ -th block diagonal matrix, where  $W_{k,ij}$  denotes the  $(i, j)$ -th element of the matrix  $\mathbf{W}_k$ . As  $\psi_k(\mathbf{\Lambda})$  is a diagonal matrix, we denote  $\psi_k(\mathbf{\Lambda}) = \text{diag}([\tilde{\lambda}_{k1}, \tilde{\lambda}_{k2}, \dots, \tilde{\lambda}_{kn}])$  for simplification. Since  $\|\psi_k(\mathbf{L})\|_2 \leq 1$ , we have  $\tilde{\lambda}_{ki} \in [-1, 1]$ . And, since  $\sum_{k=0}^{K_1-1} \|\mathbf{W}_k\|_\infty < 1$ , we have

$$\begin{aligned} \|\Gamma_0^{K_1}(\mathbf{W}, \mathbf{\Lambda})\|_2 &\leq \max_{i,r} \sum_j |\mathbf{A}_{rr}^{(ij)}| = \max_{i,r} \sum_j \left| \sum_{k=0}^{K_1-1} W_{k,ij} \tilde{\lambda}_{kr} \right| \\ &\leq \max_{i,r} \sum_j \sum_{k=0}^{K_1-1} |W_{k,ij}| < 1. \end{aligned} \quad (13)$$

Therefore, when  $t \rightarrow \infty$ , we can derive  $\text{vec}(\hat{\mathbf{Y}}_{t+1})$  as

$$\text{vec}(\hat{\mathbf{Y}}_{t+1}) = (\mathbf{I} - \Gamma_0^{K_1}(\mathbf{W}, \mathbf{\Lambda}))^{-1} \Gamma_0^1(\mathbf{V}, \mathbf{\Lambda}) \text{vec}(\hat{\mathbf{X}}_t). \quad (14)$$

After a simple algebra calculation on Eqn. (11), we can reach the conclusion of Eqn. (7)~(8).

Now we prove the bound of the transformation function  $\mathcal{T}$ . First, we can derive the following two inequations,

$$\begin{aligned} \|(\mathbf{I} - \mathbf{A}_{ii})^{-1}\|_\infty^{-1} &= \|(\mathbf{I} - \sum_{k=0}^{K_1-1} W_{k,ii} \psi_k(\mathbf{\Lambda}))^{-1}\|_\infty^{-1} \\ &= \min_l \left| 1 - \sum_{k=0}^{K_1-1} W_{k,ii} \tilde{\lambda}_{kl} \right| \geq 1 - \left| \sum_{k=0}^{K_1-1} W_{k,ii} \right|, \quad (15) \\ \sum_{j \neq i} \|\mathbf{A}_{ij}\|_\infty &= \sum_{j \neq i} \left\| \sum_{k=0}^{K_1-1} W_{k,ij} \psi_k(\mathbf{\Lambda}) \right\|_\infty \\ &= \sum_{j \neq i} \max_l \left| \sum_{k=0}^{K_1-1} W_{k,ij} \tilde{\lambda}_{kl} \right| < \sum_{j \neq i} \sum_{k=0}^{K_1-1} |W_{k,ij}|. \quad (16) \end{aligned}$$

Note the above derivation uses the conditions  $W_{k,ii} \geq 0$  and  $\tilde{\lambda}_{kl} \in [-1, 1]$ . When  $\sum_{k=0}^{K_1-1} \|\mathbf{W}_k\|_\infty < 1$ , we can further have  $\|(\mathbf{I} - \mathbf{A}_{ii})^{-1}\|_\infty^{-1} > \sum_{j \neq i} \|\mathbf{A}_{ij}\|_\infty$  for  $\forall i = 1, \dots, d'$ . So  $\mathbf{I} - \Gamma_0^{K_1}(\mathbf{W}, \mathbf{\Lambda})$  is strictly block diagonal dominant (SBDD). According to Ahlberg-Nilson-Varah bound of SSDD matrix (Morača 2007), after a series of algebra derivations, we can reach the bound:

$$\|(\mathbf{I} - \Gamma(\mathbf{W}, \mathbf{\Lambda}))^{-1}\|_\infty < \frac{1}{1 - \sum_{k=0}^{K_1} \|\mathbf{W}_k\|_\infty}. \quad (17)$$

Similarly,  $\|\Gamma_0^1(\mathbf{V}, \mathbf{\Lambda})\|_\infty \leq \|\mathbf{V}_0\|_\infty$ ,  $\|\Gamma_1^{K_2}(\mathbf{V}, \mathbf{\Lambda})\|_\infty \leq \sum_{k=1}^{K_2} \|\mathbf{V}_k\|_\infty$ . Finally we obtain an upper bound of  $\mathcal{T}$ .  $\square$

When signals are independent, *i.e.*, the recursive model takes Eqn. (4)~(5), we can have the following corollary based on the above theory.

**Proposition 1.** *Suppose the Laplacian matrix  $\mathbf{L}$  has the eigenvalue decomposition  $\mathbf{L} = \mathbf{\Phi} \text{diag}(\boldsymbol{\lambda}) \mathbf{\Phi}^\top =$*

$\mathbf{\Phi} \text{diag}([\lambda_1, \dots, \lambda_n]) \mathbf{\Phi}^\top$ , and  $\psi_k(\mathbf{L})$  is a  $k$ -order polynomial. If signals are independent, *i.e.*, taking the recursive model of Eqn. (4)~(5), for  $\forall i = 1, \dots, d, j = 1, \dots, n$ , and  $w_{ki} \geq 0$  and  $|\sum_{k=0}^{K_1-1} w_{ki} \psi_k(\lambda_j)| < 1$ , then the frequency response  $\hat{\mathbf{O}}$  in Eqn. (5) will converge to

$$\lim_{t \rightarrow \infty} \hat{\mathbf{O}}_{t+1} = \mathcal{T} \odot \hat{\mathbf{X}}_t, \quad (18)$$

$$\mathcal{T} = \frac{\mathbf{1} \mathbf{v}_0^\top}{1 - \sum_{k=0}^{K_1-1} \psi_k(\boldsymbol{\lambda}) \mathbf{w}_k^\top} + \sum_{k=1}^{K_2-1} \psi_k(\boldsymbol{\lambda}) \mathbf{v}_k^\top, \quad (19)$$

where  $-, \odot$  respectively denote the elementwise division and multiplication. Further, if  $|\psi_k(\lambda_j)| < 1$ , and considering  $\mathcal{T}$  is the elementwise multiplication, then for  $\forall (i, j)$  term of  $\mathcal{T}$ , we have an upper-bound:

$$\|\mathcal{T}_{ij}\|_2 < \frac{\|\mathbf{v}_0\|_\infty}{1 - \|\sum_{k=0}^{K_1-1} \mathbf{w}_k\|_\infty} + \sum_{k=1}^{K_2} \|\mathbf{v}_k\|_\infty. \quad (20)$$

*Proof.* According to Eqn. (12) in Theory 1, let  $\Upsilon_a^b(\mathbf{w}, \boldsymbol{\lambda}) = \sum_{k=a}^{b-1} \psi_k(\boldsymbol{\lambda}) \mathbf{w}_k^\top$ , we can have

$$\begin{aligned} \hat{\mathbf{Y}}_{t+1} &= \sum_{\tau=0}^t ((\Upsilon_0^{K_1}(\mathbf{w}, \boldsymbol{\lambda}))^\tau \odot \Upsilon_0^1(\mathbf{v}, \boldsymbol{\lambda}) \odot \hat{\mathbf{X}}_t \\ &\quad + (\Upsilon_0^{K_1}(\mathbf{w}, \boldsymbol{\lambda}))^{t+1} \odot \hat{\mathbf{Y}}_0). \end{aligned} \quad (21)$$

Since  $|\sum_{k=0}^{K_1-1} w_{ki} \psi_k(\lambda_j)| < 1$ , when  $t \rightarrow \infty$ , we can have

$$\hat{\mathbf{Y}}_{t+1} = \frac{\Upsilon_0^1(\mathbf{v}, \boldsymbol{\lambda})}{1 - \Upsilon_0^{K_1}(\mathbf{w}, \boldsymbol{\lambda})} \odot \hat{\mathbf{X}}_t. \quad (22)$$

Further, Eqn. (18) can be obtained after simple derivations. The upper-bound of  $\mathcal{T}$  can also be easily obtained.  $\square$

In practice we can take some normalization strategies on  $\{\mathbf{L}, \mathbf{W}\}$  to make them satisfy these preconditions, in order to guarantee model stability.

**Deep Stacking** The above recursive convolutional model can be easily extended into a deep architecture. Taking the recursive model as one basic layer, we may stack it into a multi-layer network architecture, in which the output signal  $\mathbf{O}$  at the bottom layer is used as the input of the top layer. With the increase of layers, the receptive field size of convolutional kernels can become larger, thus the top layer can abstract more global information. As observed from our experiments, this deep architecture (named deep STGC) can improve the performance of skeleton based action recognition. Besides, we may insert the recursive model into other networks as one basic unit to form a mixture network, according to the requirement of solved problems.

## 4 Experiments

We conduct experiments on four public skeleton based action datasets: Florence 3D (Seidenari et al. 2013), HDM05 (Müller et al. 2007), Large Scale Combined dataset (Zhang et al. 2016) and NTU RGB+D (Shahroudy et al. 2016). To investigate the effectiveness of our model, we conduct extensive experiments with different configurations listed as follows:

Table 1: Performance comparisons with different configurations for our model.

| Method                          | Florence      | HDM05         |                    | LSC           |               |               |               | NTU           |               |
|---------------------------------|---------------|---------------|--------------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                                 |               | Protocol 1    | Protocol 2         | Cross Sample  |               | Cross Subject |               | Cross View    | Cross Subject |
|                                 | Accuracy      | Accuracy      | Accuracy           | Precision     | Recall        | Precision     | Recall        | Accuracy      | Accuracy      |
| STGC (w/o L)                    | 98.14%        | 73.82%        | 81.92%±1.04        | 82.51%        | 80.80%        | 82.44%        | 81.12%        | 80.87%        | 70.29%        |
| STGC (w/ L)                     | 98.60%        | 75.76%        | 82.24%±1.18        | 83.65%        | 82.04%        | 83.63%        | 80.24%        | 81.11%        | 71.00%        |
| STGC <sub>K</sub> (indep.)      | 98.14%        | 77.08%        | 82.49%±1.49        | 84.70%        | 82.60%        | 82.08%        | 78.17%        | 79.71%        | 70.62%        |
| STGC <sub>K</sub> (dep.)        | 97.67%        | 77.99%        | 83.40%±1.30        | 86.44%        | 85.10%        | 84.00%        | 82.12%        | 81.84%        | 72.09%        |
| Deep STGC <sub>K</sub> (indep.) | 98.60%        | 78.19%        | <b>86.17%±1.25</b> | 86.73%        | 86.18%        | 84.58%        | 81.11%        | 83.57%        | 73.60%        |
| Deep STGC <sub>K</sub> (dep.)   | <b>99.07%</b> | <b>78.68%</b> | 85.29%±1.33        | <b>88.11%</b> | <b>87.03%</b> | <b>85.44%</b> | <b>83.42%</b> | <b>86.28%</b> | <b>74.85%</b> |

- STGC (w/o L) is the standard baseline without considering neighbor nodes, *i.e.*, only performing channel mapping;
- STGC w/ L uses convolutional kernels with the receptive field of one hop neighborhood;
- STGC<sub>K</sub> (indep.) is the multi-scale filtering version under the assumption of signal independency.
- STGC<sub>K</sub> (dep.) is the general multi-scale filtering version, where signals are dependent on each other.
- Deep STGC<sub>K</sub> (indep.) is the deep scheme by stacking STGC<sub>K</sub> (indep.).
- Deep STGC<sub>K</sub> (dep.) is the deep scheme by stacking STGC<sub>K</sub> (dep.).

#### 4.1 Datasets and Settings

**Florence 3D (Florence)** This dataset was collected from a stationary Kinect, and each body skeleton was recorded with only 15 joints. It contains 215 action sequences of 10 subjects with 9 actions: wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch, and bow. Due to a few skeletal joints, some types of actions are difficult to distinguish, such as drink from a bottle, answer phone and read watch. We follow the standard experimental settings to perform leave-one-subject-out cross-validation (Wang et al. 2016a).

**HDM05** This dataset was captured by using an optical marker-based Vicon system, and contains 2337 action sequences of 130 motion classes, which are acted by 5 non-professional actors named “bd”, “bk”, “dg”, “mm” and “tr”. Each skeleton data is represented with 31 joints. Until now, this dataset should involve the most skeleton-based action categories to the best of our knowledge. Due to the intra-class variations and large number of motion classes, this dataset is challenging in action recognition. To compare with those previous literatures, we conduct two types of experiments by following two widely-used protocols. Firstly, we use two subjects “bd” and “mm” for training and the remaining three for testing (Wang et al. 2015a). Secondly, to fairly compare the current deep learning methods, we conduct 10 random evaluations, each of which randomly selects half of the sequences for training and the rest for testing (Huang and Van Gool 2017).

**Large Scale Combined (LSC)** This dataset combines nine publicly available datasets (Li, Zhang, and Liu 2010; Wang et al. 2015b; 2016b; Xia, Chen, and Aggarwal 2012;

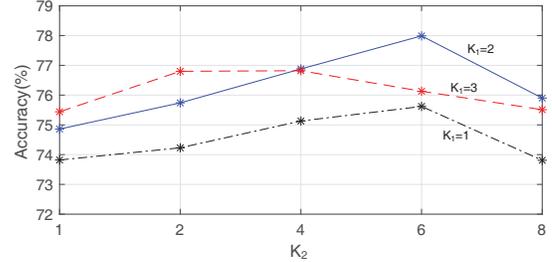


Figure 1: Comparisons of different convolutional kernel scale  $K_1, K_2$  on HDM05 (Protocol 1) for our SGTC.

Wang et al. 2012; Oreifej and Liu 2013; Koppula, Gupta, and Saxena 2013; Sung et al. 2012; Bloom, Argyriou, and Makris 2013; Bloom, Makris, and Argyriou 2012; Ni, Wang, and Moulin 2011; Chen, Jafari, and Kehtarnavaz 2015), and form a complex action dataset with 88 actions. As each individual dataset has its own characteristics in action execution manners, backgrounds, acting positions, view angles, resolutions, and sensor types, the combination of a large number of action classes makes the dataset more challenging in suffering large intra-class variation compared to each individual dataset. Following Zhang *et al.* (Zhang et al. 2016), we conduct experiments using two standard settings, *i.e.*, random cross subject evaluation and random cross sample evaluation. For each action, half of the subjects/samples are randomly selected for training while the rest for testing.

**NTU RGB+D (NTU)** This dataset is collected by Kinect v2 cameras from different views. It consists of 56880 sequences for 60 distinct actions, including various of daily actions and pair actions performed by 40 subjects. The skeleton data is represented by 25 joints. As far as we know, this dataset is currently the largest skeleton-based action recognition dataset. The large intra-class and view point variations make this dataset great challenging. Meanwhile, a large amount of samples will bring a new challenge to the current skeleton-based action recognition methods. We follow the two types of standard evaluation protocols (Shahroudy et al. 2016), *i.e.*, cross-view evaluation, cross-subject evaluation, to perform experiments.

#### 4.2 Implementation Details

**Data Processing** As skeletal data is usually captured from multi-view points and human actions are independent on the user coordinate system, we modify the origin of the coord-

Table 2: Comparisons on Florence 3D dataset.

| Method  | Accuracy      |
|---|---------------|
| Multi-part Bag-of-Poses (Seidenari et al. 2013)     | 82.00%        |
| Riemannian Manifold (Devanne et al. 2015)           | 87.04%        |
| Lie Group (Vemulapalli, Arrate, and Chellappa 2014) | 90.88%        |
| Graph-Based (Wang et al. 2016a)                     | 91.63%        |
| MIMTL (Yang et al. 2017)                            | 95.29%        |
| P-LSTM (Shahroudy et al. 2016)                      | 95.35%        |
| STGC <sub>K</sub>                                   | 97.67%        |
| Deep STGC <sub>K</sub>                              | <b>99.07%</b> |

Table 3: Comparisons on HDM05 dataset.

| Method  | Accuracy     | Accuracy           |
|---|--------------|--------------------|
| RSR-ML (Harandi, Salzmann, and Hartley 2014)        | 40.0%        | -                  |
| Cov-RP (Tuzel, Porikli, and Meer 2006)              | 58.9%        | -                  |
| Ker-RP (Wang et al. 2015a)                          | 66.2%        | -                  |
| SPDNet (Huang and Van Gool 2017)                    | -            | 61.45%±1.12        |
| Lie Group (Vemulapalli, Arrate, and Chellappa 2014) | -            | 70.26%±2.89        |
| LieNet (Huang et al. 2016)                          | -            | 75.78%±2.26        |
| P-LSTM (Shahroudy et al. 2016)                      | 70.4%        | 73.42%±2.05        |
| STGC <sub>K</sub>                                   | 78.0%        | 83.40%±1.30        |
| Deep STGC <sub>K</sub>                              | <b>78.7%</b> | <b>85.29%±1.33</b> |

dinate system as the orthocenter of joints for each frame of skeleton, *i.e.*,  $\mathcal{O} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$ , where  $\mathbf{x}_i \in \mathbb{R}^3$  is a 3D coordinate of the  $i$ -th joint,  $N$  is the number of joints. Specially, for NTU, we preprocess the joint coordinates in a way similar to Shahroudy *et al.* (Shahroudy et al. 2016). To enhance the robustness of model training, we perform data augmentation as widely used in previous deep learning literature (Liu et al. 2016; Shahroudy et al. 2016). Concretely, for each action sequence, we split the sequence into several equal sized subsequences, here 12 segments, and then pick one frame from each segment randomly to generate a large amount of training sequences. In addition, we randomly scale the skeletons by multiplying a factor in  $[0.98, 1.02]$  for the sake of the adaptive capability of scaling.

**Model Configuration** For the undirected attribute graph, we simply design edge connections according to human bones. If two joints are bridged with a bone, the edge weight is assigned to 1, otherwise 0. The signals of each node are set to its 3D coordinate. For the receptive field function, we use the simplest polynomial term, *i.e.*,  $\psi_k(\mathbf{L}) = \mathbf{L}^k$ , which represents  $k$ -hop neighbors. According to Theory 1, we normalize/clip  $\mathbf{L} = \mathbf{L}/\lambda_{max}$  ( $\lambda_{max} = 2$ ),  $\mathbf{W}_{k,ii} \geq 0$ ,  $\sum_k |\mathbf{W}_k| \mathbf{1} < \mathbf{1}$  after the gradient update at each iteration, to make sure the model stability. The outputs of recursive model are concatenated into a softmax layer for classification. In the default case, the dimension of output signals is  $d' = 32$ . For deep STGC, we empirically observe that stacking two layers is good enough to these datasets, thus we only employ two-layer network for Deep STGC. The outputs of each layer in deep STGC are 32, 64 dimensions. The most important factor of our model is the scale of convolutional kernels, which is analyzed in the next Section 4.3.

### 4.3 Selection of Convolutional Kernel Size

In the STGC<sub>K</sub> model, the parameters  $\{K_1, K_2\}$  control the receptive field sizes in temporal domain and spatial domain. When increasing  $K_1, K_2$ , the local filtering region will cover

Table 4: Comparisons on Large Scale Combined dataset.

| Method                             | Cross Sample |              | Cross Subject |              |
|------------------------------------|--------------|--------------|---------------|--------------|
|                                    | Precision    | Recall       | Precision     | Recall       |
| HON4D (Oreifej and Liu 2013)       | 84.6%        | 84.1%        | 63.1%         | 59.3%        |
| Dynamic Skeletons (Hu et al. 2015) | 85.9%        | 85.6%        | 74.5%         | 73.7%        |
| P-LSTM (Shahroudy et al. 2016)     | 84.2%        | 84.9%        | 76.3%         | 74.6%        |
| STGC <sub>K</sub>                  | 86.4%        | 85.1%        | 84.0%         | 82.1%        |
| Deep STGC <sub>K</sub>             | <b>88.1%</b> | <b>87.0%</b> | <b>85.4%</b>  | <b>83.4%</b> |

Table 5: Comparisons on NTU RGB+D dataset.

| Method   | Cross View  | Cross Subject |
|--|---|---------------|
|  | Lie Group (Vemulapalli, Arrate, and Chellappa 2014) | 52.76%        |
| Dynamic Skeletons (Hu et al. 2015)                     | 65.22%  | 60.23%        |
| HBRNN (Du, Wang, and Wang 2015)                        | 63.97%  | 59.07%        |
| LieNet (Huang et al. 2016)                             | 66.95%  | 61.37%        |
| Deep LSTM (Shahroudy et al. 2016)                      | 67.29%  | 60.69%        |
| P-LSTM (Shahroudy et al. 2016)                         | 70.27%  | 62.93%        |
| ST-LSTM (Liu et al. 2016)                              | 77.70%  | 69.20%        |
| STA-LSTM (Song et al. 2017)                            | 81.20%  | 73.40%        |
| GCA-LSTM (Liu et al. 2017)                             | 82.80%  | 74.40%        |
| Geometric Features (Zhang, Liu, and Xiao 2017)         | 82.39%  | 70.26%        |
| Clips + CNN + MTLN (Ke et al. 2017)*                   | 84.83%  | <b>79.57%</b> |
| View invariant(Raw Samples) (Liu, Liu, and Chen 2017)* | 82.56%  | 75.97%        |
| STGC <sub>K</sub>                                      | 81.84%  | 72.09%        |
| Deep STGC <sub>K</sub>                                 | <b>86.28%</b>                                       | 74.85%        |

\* They used the VGG-19/CNN model pre-trained on ImageNet after converting skeletons into images.

the farther hoping neighbors. To check the effect of different scales, we conduct an experiment on HDM05 dataset by searching the temporal kernel scale  $K_1$  in  $\{1, 2, 3\}$  and the spatial kernel scale  $K_2$  in  $\{1, 2, 4, 6, 8\}$ . Considering the continuous convolution filtering is performed along time slices like a stacking CNN, in practice we should employ a smaller  $K_1$  than  $K_2$ . The cross-comparison results are reported in Figure 1. The best performance is obtained at  $K_1 = 2$  and  $K_2 = 6$ , which are used as the default parameters. Note that,  $K = 1$  means STGC(w/o L), which convolves only on the node itself without any neighbors. If  $K_1 = 1$  and  $K_2 > 1$ , only spatial filtering is taken. Conversely,  $K_1 > 1$  and  $K_2 = 1$  for only temporal filtering. We can observe that the spatial and temporal filtering together contribute the gain for action recognition.

### 4.4 Verification of STGC Structure

For our proposed model STGC itself, we conduct a series of experiments with different configurations on four benchmark datasets. The results are summarized in Table 1. They include six configurations: STGC(w/o L), STGC(w/L), STGC<sub>K</sub>(indep.), STGC<sub>K</sub>(dep.), deep STGC<sub>K</sub>(indep.) and deep STGC<sub>K</sub>(dep.). The standard baseline should be STGC(w/o L), which doesn't use any adjacent relationship, *i.e.*, implementing a multi-channel mapping like  $1 \times 1$  convolution on images. When introducing the convolutional kernel with 1-neighborhood, *i.e.*, STGC(w/L), the performance is improved. As discussed in Section 4.3, we set  $K_1 = 2$  and  $K_2 = 6$  as default for multi-scale STGC<sub>K</sub>(indep.), STGC<sub>K</sub>(dep.), where the former is the version of independent signals while the latter is the general one. STGC<sub>K</sub>(dep.) is slightly superior to STGC<sub>K</sub>(indep.) due to the consideration of signal interaction. When extending STGC into the

deep architecture of two layers, we can achieve the best performance on all four benchmark datasets.

#### 4.5 Comparisons with State-of-the-Art

We compare state-of-the-art methods on Florence, HDM05, LSC, NTU, respectively, which are summarized in Table 2, 3, 4 and 5. As observed from these results, we have the following observations.

*The proposed spatio-temporal graph convolution method is superior to the recent graph-based method (Wang et al. 2016a) and LSTM-based methods (Shahroudy et al. 2016; Liu et al. 2016; 2017; Song et al. 2017).* As shown in Table 2, our STGC has a large improvement (more than 7%) in contrast to the graph-based work (Wang et al. 2016a). In principle, our STGC is very different from this work (Wang et al. 2016a), although graph is used for both. Our method falls into a recursively convolutional architecture, while the work (Wang et al. 2016a) follows the conventional graph kernel matching technique. Also, different from those LSTM-based methods, which only model dynamics of sequences by revising LSTM, our method absorbs the success of convolutional filtering into a recursive learning with a theoretical guarantee.

*Our proposed STGC improves the current state-of-the-art on most datasets.* On the Florence dataset, our method achieves a nearly perfect performance 99.07%. On the current largest dataset NTU, under the same recursive idea, the performance is pushed to the higher 86.28% and 74.85%, from 82.80% and 74.40% for GCA-LSTM. Recently, the CNN-based methods (Ke et al. 2017; Liu, Liu, and Chen 2017) converted skeletons into images and then employed the sophisticated CNN feature extraction techniques by using the pre-training on ImageNet (Deng et al. 2009). Even so, without the use of extra training data, our STGC is still more competitive over them.

*Deep learning based methods are more effective than those shallow learning methods.* The advanced nonlinear dynamic networks, variations of LSTM (Shahroudy et al. 2016; Liu et al. 2016; 2017; Song et al. 2017) and CNN-based models (Ke et al. 2017; Liu, Liu, and Chen 2017), largely improve the action recognition performance, due to their robust representation ability. For the conventional matrix-based descriptors (*e.g.*, covariance or its variants), although the deep manifold learning strategies (Huang et al. 2016; Huang and Van Gool 2017) are developed recently, the matrix-based representations limit their representation capability because the only second-order statistic relationship of skeletal joints is preserved, whereas first-order statistics is also informative (Ranzato and Hinton 2010).

*Different datasets have different performances.* Florence 3D is the simplest dataset with 215 sequences and 9 action classes, thus most methods obtain a higher accuracy. The most difficult dataset should be the largest dataset NTU, which consists of 56880 sequences and covers various of daily actions and pair actions. The cross subject accuracy is still less than 80% due to various entangled actions. *Cross subject is more difficult than cross view or cross sample.* The phenomenon is observed from Table 4, Table 5, and Table 3 (the left/right column w.r.t cross subject/cross sample). It is

easy to understand, in the cross subject task, more unforeseeable information exists in the testing set, compared to the other tasks.

## 5 Conclusion

We proposed a spatio-temporal graph convolution approach for combining the successes of local convolution filtering and the recursive learning power of autoregressive moving average. To locally filter on spatio-temporal structures, we introduced multi-scale graphical convolutional kernels, which were composed of the receptive field matrices defined by polynomials of adjacency matrix, and signal mappings. The multi-scale convolutional kernels were simultaneously performed on hidden states of sequences for encoding the motion variations and input state for extracting spatial graphical feature. In theory, we proved the convergence of the proposed model and provided an upper-bound. Moreover, we extended the basic model into a multi-layer deep architecture. We verified the representation ability of the proposed STGC and its deep version. We also demonstrated the improvements with STGC on four public skeletal datasets, including the current largest NTU RGB+D dataset. As a generic model, the proposed model may be generalized into many problems modeled by dynamic graphs, which will be one of our future work.

## Acknowledgement

This work was supported in part by the National Basic Research Program of China under Grant 2015CB351704 and 2014CB349303, in part by the National Natural Science Foundation of China under Grant Nos. 61772276, 61602244, 61472187, 61231002 and 91420201, and program for Changjiang Scholars.

## References

- Bloom, V.; Argyriou, V.; and Makris, D. 2013. Dynamic feature selection for online action recognition. In *Proc. HBU*, 64–76.
- Bloom, V.; Makris, D.; and Argyriou, V. 2012. G3d: A gaming action dataset and real time action recognition evaluation framework. In *Proc. CVPRW*, 7–12.
- Chaudhry, R.; Offi, F.; Kurillo, G.; Bajcsy, R.; and Vidal, R. 2013. Bio-inspired dynamic 3d discriminative skeletal features for human action recognition. In *Proc. CVPRW*, 471–478.
- Chen, C.; Jafari, R.; and Kehtarnavaz, N. 2015. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proc. ICIP*, 168–172.
- Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proc. NIPS*, 3837–3845.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 248–255.
- Devanne, M.; Wannous, H.; Berretti, S.; Pala, P.; Daoudi, M.; and Del Bimbo, A. 2015. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Trans. Cybern.* 45(7):1340–1352.
- Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proc. CVPR*, 1110–1118.

- Hannan, E. J., and Deistler, M. 2012. *The statistical theory of linear systems*.
- Harandi, M. T.; Salzmann, M.; and Hartley, R. 2014. From manifold to manifold: Geometry-aware dimensionality reduction for spd matrices. In *Proc. ECCV*, 17–32.
- Hu, J.-F.; Zheng, W.-S.; Lai, J.; and Zhang, J. 2015. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proc. CVPR*, 5344–5352.
- Huang, Z., and Van Gool, L. 2017. A riemannian network for spd matrix learning. In *Proc. AAAI*.
- Huang, Z.; Wan, C.; Probst, T.; and Van Gool, L. 2016. Deep learning on lie groups for skeleton-based action recognition. *arXiv preprint arXiv:1612.05877*.
- Jain, A.; Zamir, A. R.; Savarese, S.; and Saxena, A. 2016. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proc. CVPR*, 5308–5317.
- Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; and Boussaid, F. 2017. A new representation of skeleton sequences for 3d action recognition. In *Proc. CVPR*.
- Koppula, H. S.; Gupta, R.; and Saxena, A. 2013. Learning human activities and object affordances from rgb-d videos. *Int. J. Robot. Res.* 32(8):951–970.
- Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.
- Li, C.; Cui, Z.; Zheng, W.; Xu, C.; Ji, R.; and Yang, J. 2017. Action-attending graphic neural network. *arXiv preprint arXiv:1711.06427*.
- Li, W.; Zhang, Z.; and Liu, Z. 2010. Action recognition based on a bag of 3d points. In *Proc. CVPRW*, 9–14.
- Liu, J.; Shahroudy, A.; Xu, D.; and Wang, G. 2016. Spatio-temporal lstm with trust gates for 3d human action recognition. In *Proc. 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. *Proc. CVPR*.
- Liu, M.; Liu, H.; and Chen, C. 2017. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognit.* 68:346–362.
- Morača, N. 2007. Upper bounds for the infinity norm of the inverse of sdd and s-sdd matrices. *Journal of Computational and Applied Mathematics* 206(2):666–678.
- Müller, M.; Röder, T.; Clausen, M.; Eberhardt, B.; Krüger, B.; and Weber, A. 2007. Documentation mocap database hdm05.
- Ni, B.; Wang, G.; and Moulin, P. 2011. Rgb-d-hudaact: A color-depth video database for human daily activity recognition. In *Proc. ICCVW*, 1147–1153.
- Niepert, M.; Ahmed, M.; and Kutzkov, K. 2016. Learning convolutional neural networks for graphs. In *Proc. ICML*.
- Oreifej, O., and Liu, Z. 2013. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proc. CVPR*, 716–723.
- Pržulj, N. 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics* 23(2):e177–e183.
- Ranzato, M., and Hinton, G. E. 2010. Modeling pixel means and covariances using factorized third-order boltzmann machines. In *Proc. CVPR*, 2551–2558.
- Seidenari, L.; Varano, V.; Berretti, S.; Bimbo, A.; and Pala, P. 2013. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proc. CVPRW*, 479–485.
- Seo, Y.; Defferrard, M.; Vandergheynst, P.; and Bresson, X. 2016. Structured sequence modeling with graph convolutional recurrent networks. *arXiv preprint arXiv:1612.07659*.
- Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proc. CVPR*, 1010–1019.
- Shuman, D. I.; Narang, S. K.; Frossard, P.; Ortega, A.; and Vandergheynst, P. 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.* 30(3):83–98.
- 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 *Proc. AAAI*, 4263–4270.
- Sung, J.; Ponce, C.; Selman, B.; and Saxena, A. 2012. Unstructured human activity detection from rgb-d images. In *ICRA*, 842–849.
- Tuzel, O.; Porikli, F.; and Meer, P. 2006. Region covariance: A fast descriptor for detection and classification. In *ECCV*, 589–600.
- Vemulapalli, R.; Arrate, F.; and Chellappa, R. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *Proc. CVPR*, 588–595.
- Vishwanathan, S. V. N.; Schraudolph, N. N.; Kondor, R.; and Borgwardt, K. M. 2010. Graph kernels. *Journal of Machine Learning Research* 11(Apr):1201–1242.
- Wang, J.; Liu, Z.; Wu, Y.; and Yuan, J. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *Proc. CVPR*, 1290–1297.
- Wang, L.; Zhang, J.; Zhou, L.; Tang, C.; and Li, W. 2015a. Beyond covariance: Feature representation with nonlinear kernel matrices. In *Proc. ICCV*, 4570–4578.
- Wang, P.; Li, W.; Gao, Z.; Tang, C.; Zhang, J.; and Ogunbona, P. 2015b. Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring. In *Proc. ACM MM*, 1119–1122.
- Wang, P.; Yuan, C.; Hu, W.; Li, B.; and Zhang, Y. 2016a. Graph based skeleton motion representation and similarity measurement for action recognition. In *Proc. ECCV*, 370–385.
- Wang, P.; Li, W.; Gao, Z.; Zhang, J.; Tang, C.; and Ogunbona, P. O. 2016b. Action recognition from depth maps using deep convolutional neural networks. *IEEE Trans. Human Mach. Syst.* 46(4):498–509.
- Xia, L.; Chen, C.-C.; and Aggarwal, J. 2012. View invariant human action recognition using histograms of 3d joints. In *Proc. CVPRW*, 20–27.
- Yanardag, P., and Vishwanathan, S. 2015. Deep graph kernels. In *Proc. ACM SIGKDD*, 1365–1374.
- Yang, Y.; Deng, C.; Gao, S.; Liu, W.; Tao, D.; and Gao, X. 2017. Discriminative multi-instance multitask learning for 3d action recognition. *IEEE Trans. Multimedia* 19(3):519–529.
- Ye, M.; Zhang, Q.; Wang, L.; Zhu, J.; Yang, R.; and Gall, J. 2013. A survey on human motion analysis from depth data. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. 149–187.
- Zhang, J.; Li, W.; Wang, P.; Ogunbona, P.; Liu, S.; and Tang, C. 2016. A large scale rgb-d dataset for action recognition. In *ICPR*.
- Zhang, T.; Zheng, W.; Cui, Z.; and Li, C. 2017. Deep manifold-to-manifold transforming network for action recognition. *arXiv preprint arXiv:1705.10732*.
- Zhang, S.; Liu, X.; and Xiao, J. 2017. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *Proc. WACV*, 148–157.