Graph CNNs with Motif and Variable Temporal Block for Skeleton-based Action Recognition

Yu-Hui Wen[1,2], Lin Gao[1*], Hongbo Fu [3], Fang-Lue Zhang [4], Shihong Xia[1*]
[1]Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]School of Creative Media, City University of Hong Kong
[4]School of Engineering and Computer Science, Victoria University of Wellington
{wenyuhui,gaolin,xsh}@ict.ac.cn, hongbofu@cityu.edu.hk, fanglue.zhang@ecs.vuw.ac.nz

Abstract

Hierarchical structure and different semantic roles of joints in human skeleton convey important information for action recognition. Conventional graph convolution methods for modeling skeleton structure consider only physically connected neighbors of each joint, and the joints of the same type, thus failing to capture high-order information. In this work, we propose a novel model with motif-based graph convolution to encode hierarchical spatial structure, and a variable temporal dense block to exploit local temporal information over different ranges of human skeleton sequences. Moreover, we employ a non-local block to capture global dependencies of temporal domain in an attention mechanism. Our model achieves improvements over the state-of-the-art methods on two large-scale datasets.

1 Introduction

With the swell in popularity of real-time human motion estimation technologies, skeletal human motion data have become widely and cheaply available. Due to the demand of semantic descriptions of human motion data for the applications like medical monitoring, robot/human interaction, and action analysis for sports and security, researchers have paid more attention to skeletal human action recognition. Skeleton sequences are a kind of compact representation for human motion data, which can significantly reduce computational redundancy on semantic analysis. Compared to human action analysis from raw RGB or RGB-D video sequences, recognition on skeleton has much better performance when the background of a scene is complex (Ke et al. 2017; Kim and Reiter 2017; Zhang et al. 2017).

Conventional deep learning based methods usually structure a skeleton sequence by a time series of 2D or 3D joint coordinates or pseudo-images, which are then sent into recurrent neural networks (RNNs) or convolutional neural networks (CNNs) to predict action labels (Kim and Reiter 2017; Liu, Liu, and Chen 2017; Li et al. 2017; Song et al. 2017; Zhang et al. 2017). In fact, it is more natural to describe a skeleton as a graph with human body joints as nodes and bones as edges. Thus, in recent works, graph convolutional networks (GCNs), which generalize CNN from images to graphs of arbitrary structures, have been successfully adopted to model the skeleton data (Yan, Xiong, and Lin 2018; Tang et al. 2018). A spatial-temporal graph is firstly constructed to model the physical connection of joints and temporal edges between corresponding joints in consecutive frames (Yan, Xiong, and Lin 2018). Recently, GCNs are combined with progressive reinforcement learning to select key frames of the whole video for extracting more representative graph sequences to feed GCNs (Tang et al. 2018) .

Human body completes an action with the cooperation of various parts. The cues for understanding human actions not only lie in the relationship between spatially connected joints, but also exist in the potential dependency of disconnected joints. For example, our hands and head are physically disconnected, but their relationship is useful for recognizing the action "touch head". An extrinsic dependency has been introduced (Tang et al. 2018) for disconnected joints, but they take the relationship between all pair-wise disconnected joints with equal importance. On the other hand, human action consists of a series of important stages, indicating that there should be uneven contribution from different frames to improve the efficiency. Traditional temporal convolution methods only utilize fixed kernel sizes (Yan, Xiong, and Lin 2018), which cannot take full advantage of local temporal structure.

We propose a novel deep architecture for skeletal human action recognition by better modelling the spatial and temporal features of human actions. The basic structure is a spatial-temporal module (STM) which contains motif-based GCNs with variable temporal dense block (VTDB), as shown in Fig. 1. The motif-based graph convolution can model the dependency of physically connected and disconnected joints simultaneously, thus resulting in a more effective extraction of high-order spatial information. For the temporal modeling, VTDB is used to encode short-, mid- and long-range local information. In addition, we employ the recent non-local neural network module (Wang et al. 2018) to compute a representative sequence. This non-local block in our network is able to capture whole-range dependencies in an attention mechanism to enhance the ability for extracting global temporal features.

The major contributions of our deep learning architec-

*Corresponding Author

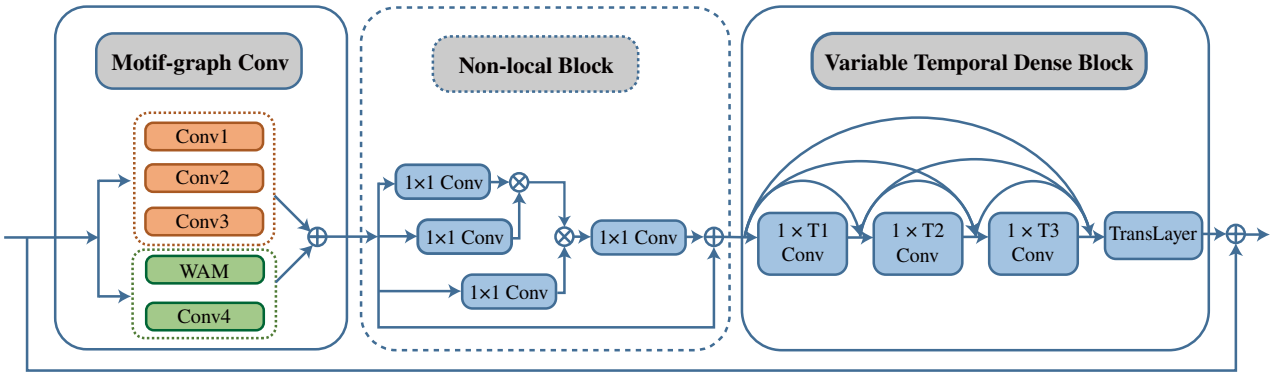

Figure 1: The architecture of the proposed spatial-temporal module (STM). It contains a motif-based graph convolution submodule for modeling spatial information, where a weighted adjacency matrix (WAM) is used for modeling action-specific spatial structure. Variable temporal dense block (VTDB) is used to encode temporal features from different ranges ($T1$, $T2$ and $T3$). TransLayer represents the transition layer in VTDB. A residual connection is applied on each STM. The non-local block is used only in the last stage of our network to reduce computation, so it is shown in a dotted line block. $\oplus$ denotes element-wise sum, and $\otimes$ denotes matrix multiplication.

ture for skeleton-based human action recognition lie in the following aspects: 1) We propose motif-based GCNs for modelling spatial information by adopting a weight-sharing scheme among joints of identical semantic roles, and constructing high-order locality by considering both physical connection and disconnection in human skeleton. 2) We propose VTDB with different kernel sizes in a dense block to capture local temporal structure over different ranges. A non-local enhanced block is incorporated into our network for constructing global dependencies in the temporal domain. 3) We propose a novel sub-structure with the combination of motif-GCNs and VTDB as a spatial-temporal module to model skeleton sequences.

We evaluate our model on two skeleton-based action recognition datasets, and the proposed model achieves improvements over previous state-of-the-art methods. Our code will be publicly available.

2 Related Work

Skeleton-based Action Recognition. With reliable skeleton data extracted by robust pose estimation algorithms from depth sensors (Shotton et al. 2011) or a single RGB camera (Cao et al. 2017; Xiu et al. 2018), skeleton-based action recognition draws more and more attention from researchers. Deep learning has been widely used in modeling the spatial and temporal patterns of skeleton sequences in this field. Many methods use RNNs due to its advantage for learning long-term sequence data (Song et al. 2017; Zhang et al. 2017). Spatio-temporal graph which models the relationship of human body components (spine, arm and leg) has also been introduced into RNNs (Jain et al. 2016), but it should be more effective to model the human parts in a finer way with every joint into the graph. CNNs has shown its superiority to RNNs owing to the parallelization over every element in a sequence and simpler training process. Skeleton sequences are manually transformed into images (Liu, Liu, and Chen 2017) to feed into CNNs (Li et al. 2017), which obtain promising performance in action recognition. Nevertheless, GCNs have shown more promising results (Yan, Xiong, and Lin 2018; Tang et al. 2018), because the images used by CNNs cannot fully describe the topology structure of skeletons. Conventional GCNs for human skeleton consider only physical connection between joints. The extrinsic dependency, referring to physically disconnected joints, which is presented with equal importance has been adopted (Tang et al. 2018). However, we take the distance between pair-wise disconnected joints into consideration to define their relationship, because close disconnected joints present significant importance for recognizing actions.

Graph Convolution. Generalizing convolutions to graph structure is achieving an increasing interest. The principles of constructing GCNs on graphs are generally categorized as spatial and spectral approaches. Spatial approaches require data preprocessing to define convolution directly on the graph nodes and their neighbors (Duvenaud et al. 2015; Atwood and Towsley 2016; Hamilton, Ying, and Leskovec 2017). Spectral approaches perform graph convolution in frequency domain (Henaff, Bruna, and LeCun 2015; Defferrard, Bresson, and Vandergheynst 2016; Kipf and Welling 2016). The Laplacian eigenbasis that transforms a graph from spatial to spectral domain should be computed and is unique for each graph structure, resulting in potentially large computation cost, non-spatially localized filters and restriction to homogeneous input. An approximation of spectral filtering is proposed by means of a Chebyshev expansion of graph Laplacian, removing intense computation and yielding spatially localized filters (Defferrard, Bresson, and Vandergheynst 2016). The idea is further simplified by limiting the filters to operate on 1-neighbor around each node (Kipf and Welling 2016). Considering high-order information in the structure of human skeleton, our network is constructed by introducing the motif notation (Sankar, Zhang, and Chen-

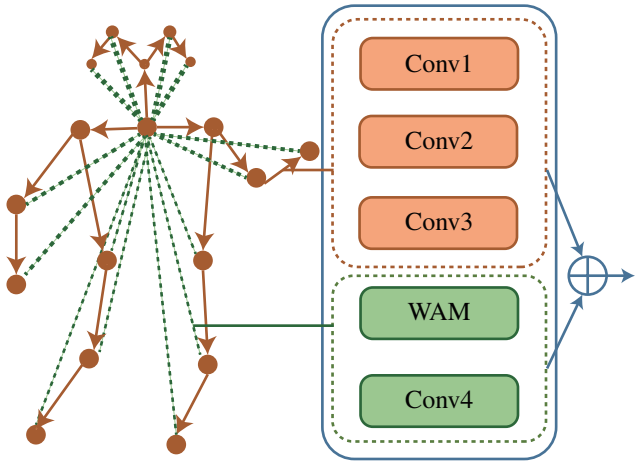

Figure 2: Spatial graph of a skeleton (on the left) and motif-based graph convolution block (on the right). In the graph, solid dots denote the body joints, while edges reflect the physical connection and disconnection with solid and dotted lines respectively. Arrows show the direction from parent nodes to child nodes. For simplicity, we only show disconnected relationship between the neck joint and other joints while all pair-wise dependency of joints are considered. $\oplus$ denotes element-wise sum.

Chuan Chang 2017) to model each joint in a skeleton with features extracted from semantically relevant joints, such as parent, child and disconnected joints.

Temporal Structure. For action video classification with RGB clips as input, 3D convolution (Yue-Hei Ng et al. 2015; Tran et al. 2015; Carreira and Zisserman 2017) is the most intuitive. Nevertheless, separating a 3D convolution into spatial and temporal components yields improvements in performance (Tran et al. 2018). This idea can also be introduced into 2D convolution for skeleton sequence as input. Although most works extract features by only considering fixed temporal kernel depths, temporal transition layer that integrates feature maps of variable temporal depths has been proposed for 3D convolution (Diba et al. 2018). Recently, a non-local neural networks have been constructed for action recognition. In a non-local operation, the response at a position is computed as a weighted sum of the features at all positions in the feature maps. When the set of positions are time points, the non-local block can be used for our skeleton sequence. We incorporate variable temporal and non-local operations into our model to extract more effective local and global temporal features.

3 Approach

Human skeleton is conventionally considered as an articulated system composed of joints and rigid bones, implying a graph structure with the joints as nodes and bones as edges. GCNs can be used to learn the spatial relationship between joints. Specifically, we construct motif-based GCNs (Sankar, Zhang, and Chen-Chuan Chang 2017) to learn the high-order spatial structure, which can not only model the dependency of physically connected and disconnected joints but also the directed structure in the human skeleton. Then, we concatenate feature maps of spatial context in sequence and feed them into VTDB. The aim of VTDB is to make use of feature maps extracted at different local temporal windows over shorter and longer time ranges. Moreover, we incorporate a non-local block for global dependencies in the temporal domain to enhance the feature representation.

3.1 Motif-based Graph Convolution

For joints and possible relationship between each pair of joints, we construct a graph $G(X, \mathcal{A})$ with a node set X and an adjacency matrix $\mathcal{A}$ for the edges. In X, there are N nodes containing 2D or 3D joint coordinates. $\mathcal{A}$ encodes the relationship between each pair of nodes.

Conventional graph-based convolution for human skeleton considers only physical connection between joints (Yan, Xiong, and Lin 2018). Although extrinsic dependency has been adopted (Tang et al. 2018), it only builds relationship between disconnected joints with a weighted adjacency matrix (WAM) that treats pair-wise extrinsic dependency between joints equally. However, the relevance between disconnected joints is different. For example, relative positions of left hand and left shoulder should be more relevant than left hand and right foot in most actions. Therefor, a reasonable way to measure the possible degree of relevance is to take spatial distance between each pair of nodes into consideration. We use the Euclidean distance between disconnected joints to define WAM for modeling action-specific spatial structure. Furthermore, we adopt motif-GCNs (Sankar, Zhang, and Chen-Chuan Chang 2017) to effectively capture high-order structural information.

A motif (or metagraph) has been traditionally defined as a pattern of connections among different node types (Benson, Gleich, and Leskovec 2016; Fang et al. 2016). Here, we build a two-motif structure to model the physical connection and disconnection of joints in a human skeleton, and its architecture is illustrated in Fig. 2. The first motif M_1 encodes physically connected joints using the immediate neighboring relationship (Kipf and Welling 2016). We define three semantic roles ($K_{M_1} = 3$) to the immediate neighbors of each joint: the joint itself, the joint's parent node and child node, which inherently lie in a directed graph. In this way, we get a hierarchical structure of human body parts in a motion. The second motif M_2 is used to encode the possible relationship between disconnected joints. We define one semantic dependency ($K_{M_2} = 1$) to represent the underlying relationship. Additionally, we define WAM for disconnected joints by assigning larger weights to joints with shorter distance, since they are more important for action recognition. In detail, $\alpha_{i,j}$ in WAM for the extrinsic relationship of node i and node j is defined as $\alpha_{i,j} = \max e - e(i,j)$, where e is a matrix representing the average euclidean distance between pair-wise nodes in a sequence. Finally, WAM is normalized to reduce the bias introduced by highly connected nodes.

For input $X_t \in \mathbf{R}^{N \times D}$ with N nodes and D dimension coordinates at frame t, inspired by previous ap-

proaches (Kipf and Welling 2016; Sankar, Zhang, and Chen-Chuan Chang 2017), we implement the motif-based graph convolution for each motif M with the following formula:

$$Z_t^M = (\mathcal{D}^M)^{-1} \sum_{k=1}^{K_M} \mathcal{A}_k^M X_t W_k^M \tag{1}$$

where K_M matrices $\mathcal{A}_k^M$ define the motif-adjacency tensor $\mathcal{A}^M$ that encodes the unique semantic roles of nodes for all instances of motif M in a graph G. $\mathcal{D}$ is a diagonal degree matrix with $\mathcal{D}_{ii}^M = \sum_{j=1}^{N} \sum_{k=1}^{K_M} \mathcal{A}_{kij}^M$. W_k^M is a tensor of filter parameters for node type k in motif M, and Z_t^M is the output of the motif. If we use only a motif with one semantic role, our motif-based GCNs reduce to traditional GCNs (Kipf and Welling 2016).

3.2 Variable Temporal Dense Block

For a video with F frames, we can first build a spatial graph with joints and edges in each frame, and then add edges for the same joints between neighboring frames to form a spatial-temporal graph (Yan, Xiong, and Lin 2018). Instead of directly applying graph-based convolution on the spatial-temporal graph, we use separate spatial and temporal sub-modules because it is easier to learn with spatio-temporal decomposition methods (Tran et al. 2018). For each graph at t-th frame, we first feed it into motif-GCNs implemented as Eq. 1 and then concatenate the output Z_t in time axis to obtain a 3D tensor, which can be sent into VTDB (Fig. 1) for action recognition.

In VTDB, instead of using only one temporal window size, we use three different time ranges (T_1, T_2 and T_3) to capture more informative temporal features from shorter and longer terms. We utilize consecutive convolution layers which are densely concatenated (Huang et al. 2017), because dense blocks are parameter-efficient and have dense knowledge propagation. Thus, the feature of a skeleton sequence can be extracted with stacked STMs as shown in Fig. 3.

Dense Connectivity. In a dense block, the l-th layer is directly connected to all preceding layers and the output h_l of the composite function H_l is denoted as:

$$h_l = H_l([h_0, h_1, ..., h_{l-1}]) \tag{2}$$

where $[h_0, h_1, ..., h_{l-1}]$ denotes the concatenation of all preceding feature-maps and $H_l(\cdot)$ is a composite function of BN-RELU-Conv operations (Huang et al. 2017). Each dense block in our network has three layers to model different local temporal ranges, and we use different kernels of size $1 \times T_1$, $1 \times T_2$ and $1 \times T_3$ for the convolution operations in the layers, respectively.

Growth Rate. If the features extracted by H_l has d channels, we should get $d_0 + d \times (l-1)$ input feature-maps for the l-th layer, where d_0 is the input channels. The hyper-parameter d is referred to as the "growth rate". The growth rate can be relatively small. Specifically, we set the growth rate of each dense block according to the number of output channels of each STM.

Transition Layer. The transition layer here is used for controlling the output channels of VTDB and facilitating down-sampling when it is necessary. A transition layer consists of a batch normalization layer, a RELU layer, and a 1×1 convolutional layer. We use a max pooling layer for down-sampling.

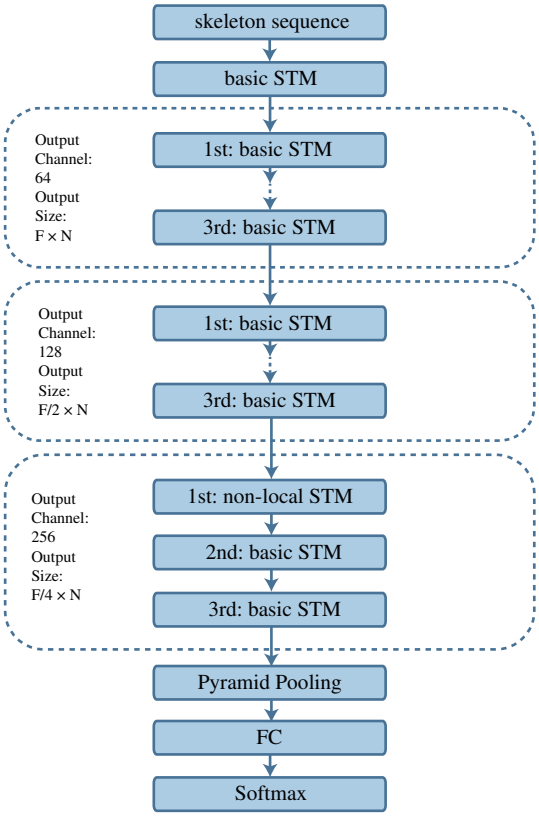

Figure 3: The architecture of our proposed motif-based graph convolution network with variable temporal dense blocks, which is composed of multiple STMs. The non-local STM is used before the last two STMs to further expand the temporal dependency.

3.3 Non-local Block for Temporal Attention

In a full skeleton sequence, every frame has different importance for modeling the temporal information. We thus use a non-local block (Wang et al. 2018), which can relate different time points of a sequence in order to compute a representation of the sequence in an attention mechanism. The features extracted by motif-GCNs denoted as $Z_{i=1...T}$, whose dimension is $T \times N \times C$, are fed into the non-local block. The non-local operation computes the dependency at a time point directly by attending at any other point in a sequence as:

$$Z_i = \frac{1}{C(Z)} \sum_{\forall j} f(Z_i, Z_j) g(Z_j) \tag{3}$$

where Z_i and Z_j denote the features extracted by spatial convolution at time i and j. $C(Z)$ is used for normalization

while $g(\cdot)$ is the unary function that computes the representation of Z_j. The pair-wise function f defines the relationship with embedded Gaussian following (Wang et al. 2018) as:

$$f(Z_i, Z_j) = e^{\theta(Z_i)^T \phi(Z_j)} \tag{4}$$

where $\theta(\cdot)$ and $\phi(\cdot)$ are the feature embedding operations with $\theta(Z_i) = W_\theta Z_i$ and $\phi(Z_j) = W_\phi Z_j$.

The non-local operation computes the response at a time point by considering the features at all temporal positions, so it is able to enlarge the receptive field from a local range to the entire sequence. In this way, we use a non-local block for temporal attention to get a more effective representation of the temporal domain, as shown in Fig. 1.

4 Experiments

We have conducted experiments on two large-scale datasets to evaluate our proposed method, and it outperforms the state-of-the-art methods on both datasets.

4.1 Network Architecture

Our network is composed of 10 layers of STM, which models spatial information with motif-based GCNs and temporal information with VTDB. Before feeding input skeleton data into our network, we use a batch normalization layer to normalize the data. The output channel numbers of STM layers are illustrated in Fig. 3. In more detail, we use separate spatial and temporal sub-modules to learn the skeleton sequence. To reduce computation, we set a smaller number of the output channels of motif graph convolution than that of STM in each layer (one half of it by default). Then, the feature maps extracted by the spatial sub-module are sent into the temporal sub-module in which the number of output channels is the same as the setting of STM. A residual connection is applied to each STM (Fig. 1).

The first motif in our 2-motif GCNs is implemented by performing three standard 1×1 2D convolution operations and multiplying the resulting tensor with a motif-adjacency matrix. In particular, we add a mask in the spatial convolution of each layer to enhance the motif's performance (Yan, Xiong, and Lin 2018). The second motif that encodes the dependency between physically disconnected joints has 1 semantic role, and we use another 1×1 2D convolution. Different from the way to model the relationship of connected joints, we use a weighted adjacency matrix in the second motif for action-specific spatial modeling.

In VTDB, we set different kernel sizes ($T_1 = 7$, $T_2 = 9$, $T_3 = 11$) for the 3 layers of a dense block. The growth rate of VTDB at each layer of STM is decided by the number of output channels. Moreover, we insert a non-local block which uses a sub-sampling strategy to reduce computation as suggested by (Wang et al. 2018) before VTDB at the 8-th STM. Max-pooling is used right after the 4-th and 7-th layers of STM for spatial sub-sampling. A pyramid pooling (He et al. 2014) is performed on the resulting feature map, which is then fed to a SoftMax classifier.

4.2 Dataset and Experiment Settings

Kinetics-M. Deepmind Kinetics (Kay et al. 2017) is an action video dataset that contains around 300,000 clips in 400 classes sourced from YouTube, and each clip lasts about 10 seconds. The dataset provides only raw video clips without skeleton data. The 2D coordinates of 18 joints (Fig. 2) have been estimated on each frame of clips (Yan, Xiong, and Lin 2018) with the realtime Openpose toolbox (Cao et al. 2017). For multi-person clips, 2 people are selected based on the average joint confidence. The dataset has been released for research purposes. However, it contains many action classes which require recognizing relationship between the actors and complicated scenes. The works focusing on skeleton-based action recognition would be inferior to video-based methods on this dataset, but a subset of 30 action classes strongly related with body motions can make the performance gap smaller (Yan, Xiong, and Lin 2018). We use the subset named as "Kinetics-M" here to evaluate our model with top-1 and top-5 classification accuracy as recommended (Kay et al. 2017). The dataset provides a training set of 25,000 clips and a validation set of 1,500 clips. The training set is used to train our models, and we report the accuracies on the validation set.

NTU-RGB+D. The widely used NTU-RGB+D is the largest dataset with annotated 3D joint coordinates for the human action recognition task (Shahroudy et al. 2016). It contains more than 56,000 sequences in 60 action classes. These clips are captured from 40 volunteers in a lab environment, where 3 camera views are recorded simultaneously from different perspectives. The provided annotations are given in the camera coordinate system. There are 25 major joints for each subject in a skeleton sequences and each sequence is guaranteed to have at most 2 subjects. The authors of this dataset recommend two benchmarks: 1) cross-subject (CS) benchmark with 40,320 and 16,560 clips for training and evaluation, respectively. In this setting, the training set comes from one subset of 20 subjects and a model is validated on sequences from the remaining 20 subjects; 2) cross-view (CV) benchmark with 37,920 clips for training and 18,960 for evaluation. Training samples in this setting come from the camera views 2 and 3, and the evaluation samples are all from the camera view 1. We follow the conventional settings and report our results on both recommended benchmarks to compare with previous state-of-the-art methods.

Training Details. Our experiments are conducted on the PyTorch deep learning framework with 2 TITANX GPUs. The models are learned using stochastic gradient descent with Nesterov momentum (0.9). Initial learning rate is set to 0.1, and is divided by 10 at 50% and 90% of the total number of training epochs. The weight decay is 0.0001. We use random moving and selecting data augmentation methods (Yan, Xiong, and Lin 2018) when training on Kinetics-M dataset with the batch size set to 64. We set the batch size according to the available GPU memory for NTU-RGB+D dataset training without data augmentation.

	Top-1	Top-5
uni-GCNs	76.4%	94.0%
motif-GCNs	82.5%	95.7%
motif-GCNs + VTDB	83.3%	95.9%
motif-GCNs + non-local block	83.2%	95.6%
motif-GCNs + non-local VTDB	**84.2%**	**96.1%**

Table 1: Ablation study on Kinetics-M dataset. The meaning of each setting is explained in Sec. 4.3.

IC	GR	Top-1
OC/4	OC/4	82.9%
OC/2	OC/4	83.3%
OC/2	OC/2	83.5%

Table 2: The accuracies of the network with different numbers of input channels (IC) and growth-rate (GR) in VTDB. IC and GR are decided by the number of output channels (OC) of spatial-temporal module at each layer.

4.3 Ablation Study

To analyze the necessity of each proposed component, we perform a series of detailed experiments on Kinetics-M dataset.

Motif-based Graph Convolution. We use the uni-labeling partition strategy in GCNs (Yan, Xiong, and Lin 2018) as our baseline, named as uni-GCNs here, because it approximates the propagation rule of traditional GCNs (Kipf and Welling 2016), which can be seen as using only 1 motif with 1 semantic role. We compare the performance of uni-GCNs with that of motif-GCNs in modeling spatial information of human skeleton. Temporal convolution is implemented with traditional 1×1 2D convolution. As shown in Table 1, motif-GCNs outperform uni-GCNs by large margin. The result verifies that motif-GCNs improve the performance of conventional GCNs through proper modeling underlying high-order structures.

Variable Temporal Dense Block. We evaluate the necessity of another important component VTDB in our network. As shown in Table 1, VTDB shows superiority over traditional temporal convolution. The result justifies the importance of extracting local temporal information from short-, mid- and long-range terms. In addition, we have done a series of tests to achieve better configuration for VTDB, as shown in Table 2. We set the number of input channels (IC) and growth-rate (GR) with respect to the number of output channels (OC) of each STM. It is obvious that VTDB can perform better as IC and GR increase. IC is in coincidence with the number of output channels of the spatial module. When IC increases, the features extracted by the spatial module are more informative. The contribution of GR is primarily due to growth in model capacity of dense block (Huang et al. 2017). We set IC and GR (IC = OC/2, GR=OC/4 by default) considering both computation complexity and accuracy.

	Top-1	Top-5
TConv (Kim and Reiter 2017)	70.8%	92.5%
STGCN (Yan, Xiong, and Lin 2018)	79.7%	94.2%
motif-GCNs+non-local VTDB	**84.2%**	**96.1%**

Table 3: Skeleton-based action recognition performance on kinetics-M dataset in terms of top-1 and top-5 accuracies.

	CS	CV
PA-LSTM (Shahroudy et al. 2016)	62.9%	70.3%
ST-LSTM (Liu et al. 2016)	69.2%	77.7%
2S RNN (Wang and Wang 2017)	71.3%	79.5%
TConv (Kim and Reiter 2017)	74.3%	83.1%
Clips+CNN+MTLN (Ke et al. 2017)	79.6%	84.8%
VI (Liu, Liu, and Chen 2017)	80.0%	87.2%
STGCN (Yan, Xiong, and Lin 2018)	81.5%	88.3%
LSTM-CNN (Li et al. 2017)	82.9%	90.1%
DPRL (Tang et al. 2018)	83.5%	89.8%
motif-GCNs+non-local VTDB	**84.2%**	**90.2%**

Table 4: Skeleton-based action recognition performance on NTU-RGB+D dataset. The accuracies are reported on both the cross-subject (CS) and cross-view (CV) benchmarks.

Non-local Block. Table 1 shows the necessity of using a non-local block for modeling global temporal information. We use only one non-local block in our network before traditional temporal convolution or VTDB. The recognition rate can be further improved by fusing the VTDB and non-local block. The results indicate that it is necessary to model both local and global dependencies in the temporal domain.

4.4 Comparison with State of the Arts

We demonstrate the effectiveness of our model on two large-scale datasets and compare with state-of-the-art methods.

On kinetics-M dataset, we compare with TConv (Kim and Reiter 2017) and ST-GCN (Yan, Xiong, and Lin 2018) in terms of top-1 and top-5 accuracies for skeleton-based action recognition. Our model outperforms the previous spatial and temporal convolution methods on this dataset as shown in Table 3.

We compare our final model with the state-of-the-art methods on NTU-RGB+D dataset. As shown in Table 4, we report both CS and CV top-1 classification accuracies following the standard practice in literature (Shahroudy et al. 2016). Our model outperforms the state-of-the-art approaches on this dataset without data augmentation as used in previous literatures (Ke et al. 2017; Kim and Reiter 2017; Yan, Xiong, and Lin 2018).

4.5 Discussions

In the ablation study, motif-based GCNs contribute most to the improvements of our model. VTDB is prone to increase the accuracy consistently with growth in both IC and GR. Particularly, there is no sign of performance degragation or overfitting. By tuning the hyper-parameters, further gains in accuracy may be obtained. We show the superiority of our model compared to uni-GCNs intuitively in Fig. 4.

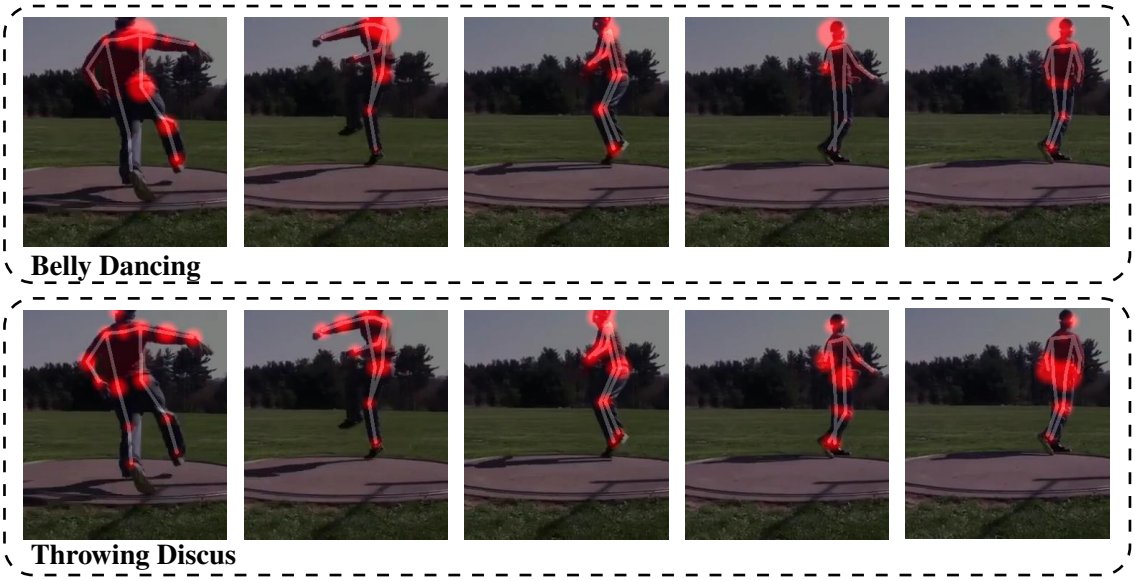

Figure 4: The response magnitude of all the joints (red dots) in a motion sequence in the last layer of uni-GCNs (the first row) and our model (the second row). Uni-GCNs gives the wrong label "Belly Dancing", while our model recognize the motion "Throwing Discus" correctly. The video sequence is selected from a clip of Kinetics-M dataset with an interval of 20 frames.

Uni-GCNs mislead "throwing discus" as "belly dancing", because the networks tend to consider only close neighbors. Our model gets more reasonable response at joints and gives the right label to the action by adopting the possible dependencies of physically disconnected joints and the hierarchical structure in the skeleton. Kinetics-M dataset is captured in unconstrained environments. Achieving improvements on it verifies that out model is able to extract more effective features from noisy data. Furthermore, we also get a competitive performance on NTU-RGB+D dataset whose data captured in constrained lab environments are more stable. Recently, there is a trend to use multi-streams of data to raise the performance of action recognition. Although we focus on skeleton-based data in this work, our model can also provide complementary information to RGB and optical flow models to raise the performance.

5 Conclusion

In this paper, we have presented a new motif-based graph convolution network with variable temporal dense block architecture for skeleton-based action recognition. The network constructs a set of spatial-temporal modules to model skeleton sequences. Motif-GCNs effectively fuse information from different semantic roles of physically connected and disconnected joints to learn high-order features. Moreover, we propose VTDB to model local temporal information from different ranges. A non-local block is combined with VTDB to further enhance the ability in modeling the global temporal dependencies. Our proposed model achieves improvements over the state-of-the-art methods on two challenging large-scale datasets.

6 Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61772499, No. 61872440, No. 61828204 and No. 61502453), Young Elite Scientists Sponsorship Program by CAST (No. 2017QNRC001), CCF-Tencent Open Fund, SenseTime Research Fund, Huawei HIRP Open Fund (No. HO2018085141), New Zealand National Science Challenge 10-Robotics Spearhead Project (Callaghan Innovation: CRS-S6-2017), the Centre for Applied Computing and Interactive Media (ACIM) of School of Creative Media, Strategic Research Grant (No. 7005176) from the City University of Hong Kong and NVIDIA Corporation with the GPU donation.

References

Atwood, J., and Towsley, D. 2016. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1993–2001.

Benson, A. R.; Gleich, D. F.; and Leskovec, J. 2016. Higher-order organization of complex networks. *Science* 353(6295):163–166.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1302–1310. IEEE.

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 4724–4733. IEEE.

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*, 3844–3852.

Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A. H.; Arzani, M. M.; Yousefzadeh, R.; and Van Gool, L. 2018. Temporal 3d convnets using temporal transition layer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1117–1121.

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

Fang, Y.; Lin, W.; Zheng, V. W.; Wu, M.; Chang, K. C.-C.; and Li, X.-L. 2016. Semantic proximity search on graphs with metagraph-based learning. In *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on*, 277–288. IEEE.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 1024–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European conference on computer vision*, 346–361. Springer.

Henaff, M.; Bruna, J.; and LeCun, Y. 2015. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, volume 1, 3.

Jain, A.; Zamir, A. R.; Savarese, S.; and Saxena, A. 2016. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5308–5317.

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 *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 4570–4579. IEEE.

Kim, T. S., and Reiter, A. 2017. Interpretable 3d human action analysis with temporal convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1623–1631. IEEE.

Kipf, T. N., and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, B.; Dai, Y.; Cheng, X.; Chen, H.; Lin, Y.; He, M.; et al. 2017. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 601–604. IEEE.

Liu, J.; Shahroudy, A.; Xu, D.; and Wang, G. 2016. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, 816–833. Springer.

Liu, M.; Liu, H.; and Chen, C. 2017. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition* 68:346–362.

Sankar, A.; Zhang, X.; and Chen-Chuan Chang, K. 2017. Motif-based convolutional neural network on graphs. *arXiv preprint arXiv:1711.05697*.

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*, 1010–1019.

Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; and Blake, A. 2011. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1297–1304. Ieee.

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 *AAAI*, volume 1, 4263–4270.

Tang, Y.; Tian, Y.; Lu, J.; Li, P.; and Zhou, J. 2018. Deep progressive reinforcement learning for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5323–5332.

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*, 4489–4497.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6450–6459.

Wang, H., and Wang, L. 2017. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *e Conference on Computer Vision and Pa ern Recognition (CVPR)*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; and Lu, C. 2018. Pose Flow: Efficient online pose tracking. *ArXiv e-prints*.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455*.

Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; and Toderici, G. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4694–4702.

Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N.; et al. 2017. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. *arXiv, no. Mar*.