Multi-level Temporal-Guided Graph Convolutional Networks for Skeleton-Based Action Recognition

Kunlun Wu(✉) and Xun Gong

School of Computing and Artificial Intelligence, Southwest Jiaotong University,
Chengdu, China
wukunlun@my.swjtu.edu.cn, xgong@swjtu.edu.cn

Abstract. Skeleton-based action recognition is a crucial and challenging task, which has promoted remarkable progress in diverse fields. Nevertheless, how to capture long-range temporal relationships remains a challenging problem, which is vital to reducing the ambiguity of indistinguishable actions. Towards this end, we propose a novel Multi-Level Temporal-Guided Graph Convolutional Network (ML-TGCN) to tackle the above problem. We leverage the multi-level temporal-guided mechanism to learn diverse temporal receptive fields for mining the discriminative motion patterns. Moreover, most current approaches cannot effectively explore the comprehensive spatial topology due to the skeleton graph is heuristically predefined, thus we propose a cross-space GCN to capture global context and maintain strengths of GCNs (i.e., hierarchy and local topology) jointly beyond the physical connectivity. The experimental results on the challenging datasets *NTU RGB+D* and *Kinetics-Skeleton* verify that ML-TGCN can achieve state-of-the-art performance.

Keywords: Skeleton-based action recognition · Multi-level temporal-guided · Graph convolutional network

1 Introduction

With the prosperity achieved in deep learning and computer vision, action recognition has accomplished better development in recent years and already applicated in various fields, such as human-computer interaction, eldercare, video surveillance and healthcare assistance. Current action recognition baselines can be categorized into video-based and skeleton-based. Rapid developments in 3D depth cameras such as Microsoft Kinect and Intel RealSense sensors, besides, human pose estimation algorithms make it more convenient to obtain 2D or 3D skeleton coordinates quickly and accurately. The skeleton-based action recognition methods have received more attention for their excellent topology-based representation and robustness to the environmental changes. Nowadays, the most dominative method to achieve skeleton-based action recognition has become Graph Neural Networks (GNNs), especially, GCNs have been investigated to be

W. Deng et al. (Eds.): CCBR 2022, LNCS 13628, pp. 267–275, 2022.
https://doi.org/10.1007/978-3-031-20233-9_27

very effective in modeling non-Euclidean data. ST-GCN [1] is the first model that uses the spatial-temporal skeleton graph for action recognition. Latter, diverse variants based on ST-GCN have boosted the recognition performance. Despite the fact that ST-GCNs have made remarkable progress, the structural constraints have limited the expressive ability of GCNs.

For ST-GCNs [1–3], the topology of the spatial graph represents the physical connection and is pre-designed for all layers, which can hinder the expressive ability due to the limited spatial-temporal receptive field. Particularly, message propagation can only flow along a fixed path when graph links are directed. Numerous studies have shown that the relationship between body joints not associated in the spatial graph is still crucial for recognition, such as "dancing" that left hand is apart from right hand. To better guide the encoder about where and when to focus on jointly, we attempt to overcome the aforementioned limitations by introducing a novel Multi-Level Temporal-Guided Graph Convolutional Network to jointly learn discriminative local-global spatiotemporal features. Intuitively, diverse temporal levels are determined by the size of the corresponding temporal segment, which can allow the model to learn fine-grained features of highly similar relations and effectively reduce the ambiguity of indistinguishable actions. Furthermore, the proposed cross-space GCN learns global context and local information by capturing the dependencies of non-local joints in the spatial graph. To summarize, the main contributions of this work lie in three folds:

- We propose a novel Multi-Level Temporal-Guided Mechanism (ML-TGM) to capture diverse temporal receptive fields, which can significantly improve recognition performance on hard samples, i.e., reducing the ambiguity of highly similar actions.
- With the aim to capture the optimal spatial topology, we develop an effective cross-space GCN to learn global context and local information simultaneously, allowing the model can mine relationships between the joints that are far away from each other in the anatomy-based graph.
- The experimental results on two large-scale datasets *NTU RGB+D* [4] and *Kinetics-Skeleton* [1] indicate our ML-TGCN can exceed state-of-the-art.

2 Methodology

In this section, we mainly illustrate the proposed method in two core parts. First, we present the proposed ML-TGM in detail. Subsequently, we analyze the drawbacks of traditional GCNs and introduce the principle of cross-space GCN.

2.1 Multi-level Temporal-Guided Mechanism

As previously emphasized in Sect. 1, understanding long-term temporal dependencies is vital for efficient relations modeling, especially for indistinguishable actions. Therefore, we propose a Multi-level Temporal-guided Mechanism (ML-TGM) to accomplish the above goal. As shown in Fig. 1, we first use a hierarchical backbone, i.e., several layers of regular ST-GCN, to obtain the general

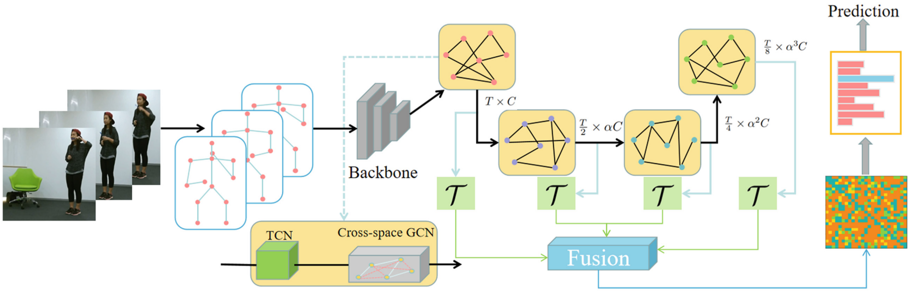

Fig. 1. Conceptual diagram of our ML-TGCN. The skeleton-based features are fed into multi-level temporal-guided GCN to extract long-range spatial-temporal relationships. The module of each level consists of the temporal convolution and the proposed cross-space GCN. Eventually, the fused multi-level features are utilized for action recognition.

features representation $\mathbf{F}_g \in \mathbb{R}^{T \times C}$. Then, the model learns multi-level temporal dependencies via the proposed mechanism. Each level consists of a temporal convolution and $\times K$ cross-space GCN, which corresponds to the different receptive fields. Earlier levels mine the fine-grained behaviour features with more temporal series, whereas the latter levels learn a coarse representation with fewer series, we leverage the interaction of different levels to guide our model for understanding long-range temporal relations. Moreover, to better model network hierarchy, temporal features merging is the essential one, thus we adopt a transformation function $\mathcal{T}$ to ensure the representation of different levels can obtain a unified feature representation $\mathbf{F}_u \in \mathbb{R}^{T \times C_m}$, which attempts to resample the series across temporal dimensions. Specifically, we accomplish it by utilizing an interpolate function and a linear mapping to the richer semantic space. We denotes the representation of each level as $\mathbf{F}^{(n)} \in \mathbb{R}^{\frac{T}{2^{n-1}} \times \alpha^{n-1} C}$, α is the channel-wise scale factor. The transformed features $\mathbf{F}_{\mathcal{T}}^{(n)} \in \mathbb{R}^{T \times C_m}$ via the function $\mathcal{T}$ can be denoted as:

$$\mathbf{F}_{\mathcal{T}}^{(n)} = Interpolation(\mathbf{F}^{(n)} \Theta^{(n)}) \tag{1}$$

$\Theta^{(n)}$ is the feature transformation at each level. Intuitively, earlier levels have lower semantics, whereas the latter levels have higher semantics. To balance the interaction of them, we adopt a trainable weight $\lambda^{(n)}$ at each level for exploring the appropriate relationships. The transformed features can be described as:

$$\tilde{\mathbf{F}}_{\mathcal{T}}^{(n)} = \lambda^{(n)} \mathbf{F}_{\mathcal{T}}^{(n)} \tag{2}$$

Here, the temporal length of all the refined representations is the same. Eventually, we merge the features at each level along the channel dimension to get the multi-level temporal representation, which can be formulated as follows:

$$\tilde{\mathbf{F}}_{\mathcal{T}} = \tilde{\mathbf{F}}_{\mathcal{T}}^{(1)} \oplus \tilde{\mathbf{F}}_{\mathcal{T}}^{(2)} \oplus ... \oplus \tilde{\mathbf{F}}_{\mathcal{T}}^{(n)} \tag{3}$$

$\oplus$ represents the channel-wise concatenation. The refined features $\tilde{\mathbf{F}}_{\mathcal{T}}$ contain diverse temporal receptive fields, which can benefit the more comprehensive long-range temporal modeling and boost the model to distinguish highly similar actions more accurately.

2.2 Cross-space GCN

Generally, the skeleton graph is described as $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ is composed of N joints, $\mathcal{E}$ represents the edge set accomplished by a symmetric adjacency matrix $A \in \mathbb{R}^{N \times N}$. The action is represented as graph sequences $X \in \mathbb{R}^{T \times N \times C}$, and the original input is defined by features X and graph structure A jointly. Existing GCN-based approaches mostly model the skeleton data by a spatial GCN and temporal convolution, which typically introduce incremental modules to enhance the expressiveness ability. However, the formed spatial-temporal receptive fields are pre-defined heuristically and distant nodes have weak information interaction. Therefore, we attempt to merge multi-scale structural features to learn higher-order polynomials of the adjacency matrix for mining global relations. The regular multi-scale formulation can be denoted as:

$$X' = \sum_{k=0}^{K} \widehat{D}^{(k)^{-\frac{1}{2}}} \widehat{A}^{(k)} \widehat{D}^{(k)^{-\frac{1}{2}}} XW \tag{4}$$

where K is the number of scales for aggregation, $\widehat{A}^{(k)}$ is the adjacency matrix of $\mathbf{A}$ at k^{th} scale, $D^{(k)}$ is the diagonal degree matrix of $\widehat{A}^{(k)}$ and W is the trainable linear transformation. Many investigations have verified that ordinary multi-scale mechanisms still concentrate more on the local region due to cyclic walks. The self-loops can create more space for cycles, resulting in the bias of the receptive field. To solve the above problem, we modify the above formulation to suppress redundant dependencies. Specifically, we propose a mechanism ϕ to reformulate the adjacency matrix at each scale, which can be formulated as:

$$\widehat{A}^{(k)} = \mu \widehat{A}^{(k)} + (\mu - 1)\widehat{A}^{(k-1)} + I \tag{5}$$

$\mu \in [0, 1]$ is a learnable parameter and I is the identity matrix, self-loops I is essential for learning the k-hop relationships and accelerating convergence.

In real scenarios, the performed actions always have complex cross-space connectivity, thus we attempt to make a more exhaustive feature interaction. Specifically, current methods treat the spatial aggregation as the fusion along the channel dimension, i.e., each channel shares the same human topology, but actually different channels have independent spatial context and should have the trainable adjacency matrix respectively. Therefore, we further modify graph convolution along the channel dimension. To be specific, we split channels into G groups, each channel in a group shares the joint-level learnable adjacency matrix. The model refines features by the operation Φ, which can be denoted as:

$$\widehat{F} = \psi(\widehat{A}^c_{1,:,:} F_{:\lfloor \frac{C}{G} \rfloor,:} || \widehat{A}^c_{2,:,:} F_{\lfloor \frac{C}{G} \rfloor : \lfloor \frac{2C}{G} \rfloor,:} || ... || \widehat{A}^c_{i,:,:} F_{\lfloor \frac{(i-1)C}{G} \rfloor : \lfloor \frac{iC}{G} \rfloor,:}) \qquad 1 \leq i \leq G \tag{6}$$

where $F \in \mathbb{R}^{C \times N}$ is the transformed feature by using the mechanism ϕ, $\widehat{A}^c \in \mathbb{R}^{G \times N \times N}$ is the grouped adjacency matrix, $||$ and ψ is the channel-wise concatenation and shuffle operation. The combination of ϕ and Φ forms the proposed cross-space GCN, with the aim to enhance the features interaction of different joints and explore the optimal spatial topology beyond the physical connectivity.

3 Experiments

3.1 Datasets

NTU RGB+D. *NTU RGB+D* is the widely used and extensive dataset of 3D joint coordinates. It contains 56880 human action video sequences in 60 classes. The publishers of *NTU RGB+D* recommend two benchmarks: Cross-Subject (*X-Sub*) and Cross-View (*X-View*). This dataset is composed of two benchmarks: 1) Cross-subject (*X-Sub*): The volunteers of each subset perform 40320 actions for training, and the complement subset contains 16560 clips for evaluation. 2) Cross-view (*X-View*): This benchmark includes 37920 and 18960 clips for forming the train and evaluation set respectively. These videos are captured by three *Kinetic* depth sensors of equal height but different viewpoints. Each skeleton graph of *NTU RGB+D* consists of 25 body key points denoted by 3D coordinates.

Kinetics-Skeleton. *Kinetic-Skeleton* dataset contains about 300,000 video clips in 400 classes collected from the Internet. The captured skeleton information contains 18 body joints, along with their 2D coordinates and confidence score. Different from *NTU RGB+D*, skeleton sequences are not provided by the depth cameras but estimated by the publicly available *OpenPose* toolbox. There are 240,436 samples for training and 19794 samples for testing. Following the conventional evaluation method, Top-1 and Top-5 accuracies are reported.

3.2 Implementation Details

All experiments are conducted on four RTX 3080 TI GPUs with the PyTorch deep learning framework. We trained our models for a total of 140 epochs with batch size 32 and SGD as optimizer on *NTU RGB+D*, while on *Kinetics-Skeleton* we trained our models for a total of 80 epochs, with batch size 128. The learning rate is set to 0.1 at the beginning and then reduced by a weight decay of 10 at the epochs 60, 90, 120 and 45, 55, 70 for *NTU RGB+D* and *Kinetics-Skeleton* respectively. Moreover, we preprocessed the data with the same procedure used in [3]. In all of these experiments, we use the standard cross-entropy loss for optimization.

3.3 Ablation Studies

We analyze the proposed module by experiments on the *X-View* benchmark of *NTU RGB+D* dataset. The Top-1 accuracy of classification is used as the evaluation criterion. For this ablation study, we verify the performance of ML-TGM and cross-space GCN. Moreover, we visualize the learned skeleton graph and corresponding adjacency matrix for a more convincing explanation.

The Effectiveness of ML-TGM. Here we focus on verifying the benefits of applying the proposed ML-TGM. From Fig. 2, we can see the accuracy of ML-TGM with different levels. The proposed ML-TGM can be viewed as the general temporal modeling when we only use a single level, and the experimental results indicate that the recognition performance can actually obtain the improvement when we utilize ML-TGM. Based on the intuition of the effectiveness and efficiency, we adopt 4 levels in our ML-TGCN. Furthermore, we make a comparison of the strong baseline (left) and ML-TGM (right) to verify the performance on the hard classes. As shown in Fig. 3, we visualize the normalized confusion matrix to show the accuracy of each hard class, and especially use red rectangles to mark classes with significant improvement, which indicates that ML-TGM can reduce the ambiguity of highly similar actions (*reading and writing*, etc.) indeed.

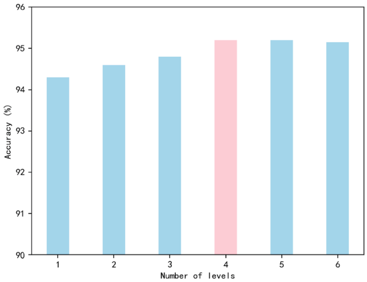

Fig. 2. The effectiveness of ML-TGM with different levels.

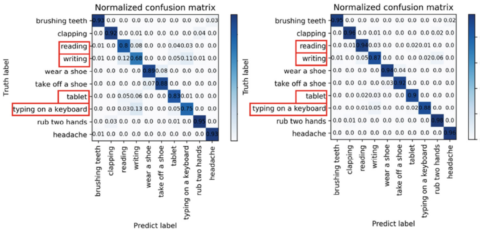

Fig. 3. The normalized confusion matrix of the strong baseline and ML-TGCN.

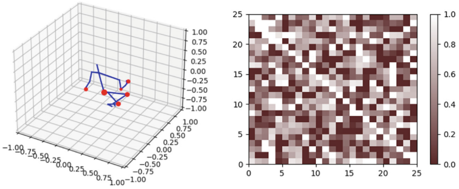

Fig. 4. The learned skeleton graph and corresponding adjacency matrix.

Table 1. The comparison of cross-space GCN and existing GCNs.

Method	Accuracy (%)
ST-GCN [1]	93.6
2s AGCN [3]	94.1
2s AGCN (Non-Local) [3]	94.4
MS-G3D [5]	94.9
Cross-space GCN	**95.2**$^{\uparrow 1.6}$

The Performance of Cross-Space GCN. To analyze the region of space that the model focuses on when performing actions, we visualize the learned skeleton graph and corresponding adjacency matrix. As shown in Fig. 4, we randomly select a test sample of "fall down" class in *NTU RGB+D* dataset, and the learned skeleton graph (left) shows that the model pays more attention on hands, knees and hips, which indicates that our model focuses actually on the action-related region. Moreover, in the learned adjacency matrix (right), we can also observe that the model actually learns global relationships, e.g., the feet and hands are apart from each other, whereas the corresponding value of the adjacency matrix is non-zero. The above result indicates that cross-space GCN captures global context beyond the physical connectivity indeed. To further verify the

Table 2. The comparisons with our ML-TGCN on *NTU RGB+D* dataset.

Datasets	Approaches	Top-1 (%)
NTU RGB+D (X-Sub)	ST-GCN [1]	81.5
	STGR-GCN [6]	86.9
	AS-GCN [2]	86.8
	2s-AGCN [3]	88.5
	DGNN [7]	89.9
	MS-G3D [5]	91.5
	Ta-CNN++ [8]	90.7
	SMotif-GCN+TBs [9]	90.5
	ML-TGCN	**91.5**
NTU RGB+D (X-View)	ST-GCN [1]	88.3
	STGR-GCN [6]	92.3
	AS-GCN [2]	94.2
	2s-AGCN [3]	95.1
	DGNN [7]	96.2
	MS-G3D [5]	95.2
	Ta-CNN++ [8]	95.1
	SMotif-GCN+TBs [9]	96.1
	ML-TGCN	**96.6**

Table 3. The comparisons with our ML-TGCN on *Kinetic-Skeleton* dataset.

Datasets	Approaches	Top-1 (%)	Top-5 (%)
Kinetic-Skeleton	ST-GCN [1]	30.7	52.8
	AS-GCN [2]	34.8	56.8
	2s-AGCN [3]	36.1	58.7
	DGNN [7]	36.9	59.6
	MS-G3D [5]	38.0	60.9
	MST-GCN [10]	38.1	60.8
	Hyper-GNN [11]	37.1	60.0
	SMotif-GCN+TBs [9]	37.8	60.6
	ML-TGCN	**38.5**	**61.2**

superiority of cross-space GCN, we make comparisons with other state-of-the-art methods and adopt the same temporal model for a fair comparison. From Table. 1, we can observe that cross-space outperforms the baseline 1.6% and have a competitive performance compared with other state-of-the-art methods.

3.4 Comparison with the State-of-the-Art

We compare our method with state-of-the-art algorithms on *NTU RGB+D (X-View)* and *NTU RGB+D (X-Sub)* respectively to verify the excellent performance of our proposed ML-TGCN. As shown in Table. 2, we report the Top-1 accuracy of these methods on both cross-subject and cross-view benchmarks of the *NTU RGB+D* dataset. For *NTU RGB+D (X-Sub)*, we can observe that ML-TGCN has a competitive performance compared with state-of-the-art methods. For example, ML-TGCN outperforms Ta-CNN++ [8] 0.8% and MST-GCN [10] 1.6% respectively. For *NTU RGB+D (X-View)*, we can also see a clear superiority as reported previously. For *Kinetic-Skeleton*, the same as current state-of-the-art methods, we leverage Top-1 and Top-5 accuracy as our evaluation metrics. As shown in Table. 3, ML-TGCN also has an obvious improvement in recognition accuracy, e.g., our model has a 0.8% performance gain compared with [9]. The above experimental results demonstrate the superiority of our ML-TGCN.

4 Conclusions

In this work, we innovatively present a Multi-Level Temporal-Guided Mechanism (ML-TGM) to capture long-range temporal relationships, which can significantly improve the recognition accuracy of indistinguishable actions, i.e., the proposed model can effectively reduce the ambiguity of highly similar actions. Moreover, we propose a cross-space GCN to capture the global context and enhance local information jointly beyond physical connectivity, with the aim to explore the optimal spatial topology. The combination of them forms a novel network called Multi-Level Temporal-Guided Graph Convolutional Network (ML-TGCN). Experimental results on two challenging datasets *NTU RGB+D* and *Kinetics-Skeleton* indicate our approach can achieve the known state-of-the-art.

References

1. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
2. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3595–3603 (2019)
3. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019) 12026–12035

4. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: a large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
5. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 143–152 (2020)
6. Li, B., Li, X., Zhang, Z., Wu, F.: Spatio-temporal graph routing for skeleton-based action recognition. Proc. AAAI Conf. Artif. Intell. **33**, 8561–8568 (2019)
7. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7912–7921 (2019)
8. Xu, K., Ye, F., Zhong, Q., Xie, D.: Topology-aware convolutional neural network for efficient skeleton-based action recognition. Proc. AAAI Conf. Artif. Intell. **36**, 2866–2874 (2022)
9. Wen, Y.H., Gao, L., Fu, H., Zhang, F.L., Xia, S., Liu, Y.J.: Motif-GCNs with local and non-local temporal blocks for skeleton-based action recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
10. Chen, Z., Li, S., Yang, B., Li, Q., Liu, H.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. Proc. AAAI Conf. Artif. Intell. **35**, 1113–1122 (2021)
11. Hao, X., Li, J., Guo, Y., Jiang, T., Yu, M.: Hypergraph neural network for skeleton-based action recognition. IEEE Trans. Image Process. **30**, 2263–2275 (2021)