Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230871

SGC-GCN: Constructing Stronger Feature Fusion Ability Graph Convolution for Skeleton-Based Action Recognition

Qibin HU, Xiaotian WANG and Qian WANG[1]
School of Electronic and Electrical Engineering, Wuhan Textile University, China

Abstract. Skeleton-based action recognition algorithms have made extensive use of graph topologies to model the connections between human skeletal joints. Constructing graph topology models with greater representational power is key to obtaining powerful feature extractors. However, existing methods are not as effective in achieving stronger modelling of associations between physically connected joints in the human skeleton. To tackle these concerns, we put forward an innovative approach involving a distinct graph topology - a graph convolutional neural network (SGC-GCN) that amalgamates a dynamic, refined graph topology with a static, shared graph topology. The dynamic partial feature extractor (CTR-GC) and the static partial feature extractor (SGC-GC) are used in conjunction with each other to obtain a stronger feature aggregation capability in the form of static reinforcement of dynamic weaker features and to achieve the objective of strengthening the correlation modelling between interconnected joints within the human skeletal structure. The combined use of the two also introduces only a small number of additional parameters, ensuring that the refined features are not affected by the noise of statically shared features. Combining SGC-GC with the temporal modelling module has resulted in the development of SGC-GCN, a graphical convolutional network with even greater feature aggregation capability. Our network demonstrates a remarkable performance surpassing existing advanced methods on the dataset(NTU RGB+D), yielding substantial advancements in action recognition accuracy.

Keywords. graph topologies; Action recognition; GCN; clusters

1. Introduction

Existing GCN-based methods have the following limitations. (1) Yan [1] et al. introduced an artificially defined human skeleton graph topology that primarily focuses on modeling the relationships among physically connected joints, neglecting the consideration of relationships between distant joint nodes. Nonetheless, in the context of human skeleton action recognition, it is crucial to capture both the connections between adjacent connected joint nodes and the connections between distant joint nodes. (2) Several recent approaches have worked on the problem of modelling relationships between distant joint nodes. As an illustration, Shi [2], Zhang [3], and Ye [4] have undertaken research focusing on dynamically acquiring the topology of the human skeleton by employing attentional mechanisms or alternative techniques. However, their approach of employing topology modeling for all channels to overcome the challenge of

[1] Corresponding Author. Qian WANG, Wuhan Textile University, China; Email: wqian@wtu.edu.cn

modeling relationships between unnaturally connected joints inadvertently results in the aggregation of features belonging to the same topology across channels. Consequently, this aggregation imposes a dual constraint, further limiting the flexibility of feature extraction. In contrast to existing approaches, Cheng [5] et al. devised a unique strategy by establishing distinct parametric topologies for each channel. Chen [6] et al. proposed an alternative way to efficiently model channel topology, the channel topology refinement convolution-CTR-GCN, which is a refinement learning It is a fine-grained learning approach to learning channel topology. By leveraging channel-specific correlations, we engage in a novel approach to refine the topology, thereby acquiring channel-specific topology tailored to the distinctive characteristics of each channel's features. This methodology not only eliminates the need for modeling the topology of each channel in isolation but also effectively mitigates the introduction of excessive parameters. However, this refinement process inadvertently overlooks the inherent relationship between the shared topology and the refined topology, as well as the intensified correlation between the static topology and the physically interconnected joints.

Based on the above limitations, in this article, our approach introduces a novel and effective fusion of static and dynamic interplay, enabling the modeling of channel topological graph convolution with exceptional accuracy and performance. SGC-GC, although using a static topology for all channels, does not limit the flexibility of feature extraction but rather serves as a novel mechanism to compensate for the effective information lost in the CTR-GC[6] feature extraction and aggregation process and does not affect the process of aggregating features in the refined topology of CTR-GC, i.e. it retains the aggregation capability of CTR-GC for refined high-level features, fills in the gaps for weaker features, and strengthens the correlation between shared and refined topologies. Specifically, our compensation mechanism topologies all channels fuse all information with the high-level features extracted by CTR-GC and augment the weaker high-level features. That is, it improves the model's ability to aggregate features, resulting in better results.

2. Structure and Mathematical Derivation Process of SGC-GCN

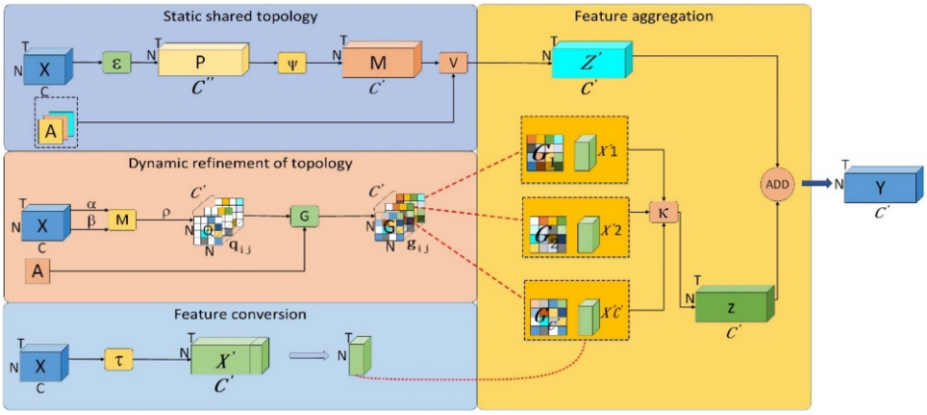

Figure 1. A framework for the interplay of static and dynamic use of graph convolution

2.1. Channel Dynamic Refinement and Static Sharing Combined with Topology Map Convolution

Input features. For a given input variable $\mathcal{X} \in \mathcal{R}^{T\times N\times C}$, the output data $z \in \mathcal{R}^{T\times N\times C}$ is formulated as:

$$z = \mathcal{A}(\mathcal{T}(\mathcal{X}), R(M(\mathcal{X}), A)) \tag{1}$$

where **A** is the learnable static shared topology.

The output data of all its channels are joined together to acquire the output data z'. The formula is expressed as:

$$\mathcal{M} = \psi(P) = \varrho \cdot P \tag{2}$$

$$z' = \mathcal{V}(\mathcal{M}, \mathcal{A}) = [\mathcal{M}1 \cdot \mathcal{A} \parallel \mathcal{M}2 \cdot \mathcal{A} \parallel \cdots \parallel \mathcal{M}\mathrm{C}' \cdot \mathcal{A}] \tag{3}$$

Feature aggregation. The yellow box in Figure 1 indicates the feature aggregation process. It constructs a channel topology map for each channel of the dynamic channel refinement topology $\mathcal{G}$ and $\mathcal{X}'$ generated by feature conversion and aggregates them in the form of channels. After concatenating the aggregated high-level features of each channel, the output z is obtained and then summed with the output z' generated by the static shared topology convolution to obtain the final output Y. The formula is as follows:

$$z = \mathrm{K}(\mathcal{X}', \mathcal{G}) = [\mathcal{X}'1 \cdot \mathcal{G}_1 \parallel \mathcal{X}'2 \cdot \mathcal{G}_2 \parallel \cdots \parallel \mathcal{X}'\mathrm{C}' \cdot \mathcal{G}_{\mathrm{C}'}] \tag{4}$$

$$\mathrm{Y} = z + z' \tag{5}$$

2.2. Model Architecture

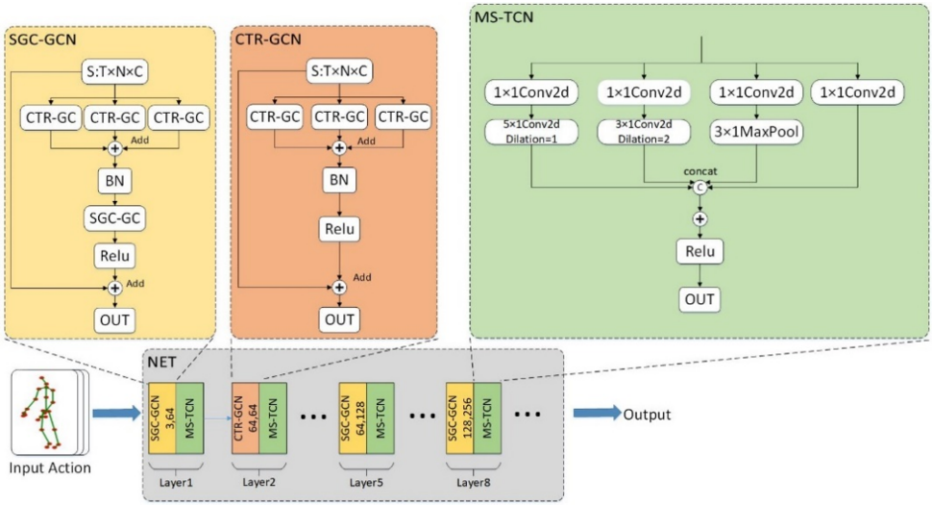

Figure 2. SGC-GCN network model diagram

The gray box in Figure 2 represents the processing process of the network for features. SGC-GCN consists of 10 submodules, each with a GCN and a TCN, which process inputs and are connected through a full connection layer. SGC-GCN was used for time dimension generation at the first, fifth and eighth layers, and CTR-GCN was used for spatial feature extraction at the other layers. The number of submodule channels ranges from 3 to 256, and the time dimensions of layers 5 and 8 are halved using step-up time convolution.

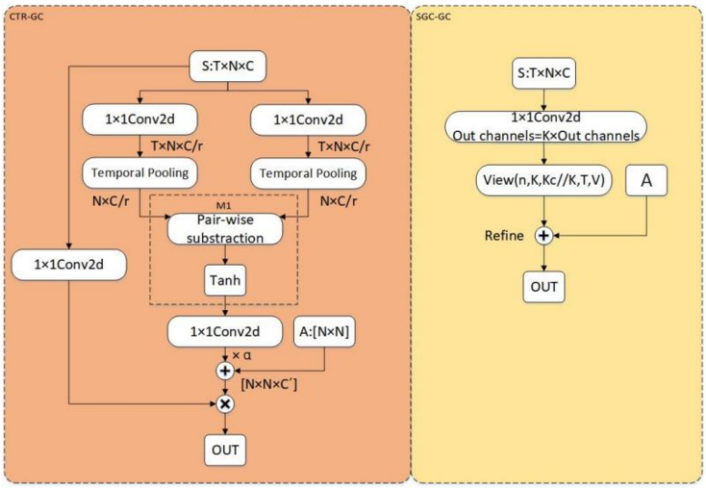

Figure 3. Feature Extractors in SGC-GCN Networks

The feature extractors used in the SGC-GCN network are CTR-GC and SGC-GC. The orange box in Figure 3 represents the feature extraction process of CTR-GC. CTR-GC performs feature extraction by compressing and aggregating along the time dimension to obtain input features, and then infers channel topology. It uses compression with a compression rate of r to obtain a compact representation of input features and performs temporal pooling. Then, specific correlation of features within each channel is obtained through bidirectional subtraction and activation. Channel topology is obtained through refined shared topology A, and feature aggregation is performed in the skeletal graph. The yellow box in Figure 3 represents the feature extraction process of SGC-GC. In SGC-GC, we reshape the matrix by adjusting the dimension of the convolution uplift matrix and applying the manually defined parameter K. This allows all channels of the input sample features to be used with the manually defined shared topology A at once, achieving matching with the temporal dimension of the dynamic part.

3. Experiments

3.1. Ablation Study

Within this section, we delve into an in-depth exploration of our novel proposition that synergistically integrates a static topology with a dynamic topology. Our primary objective is to amplify the efficacy of graph convolution and optimize its configuration on the X-sub benchmark of the NTU RGB+D dataset.

Table 1. Gradually delete or add precision comparisons in the global when using SGC-GC.

Methods	Param	X-sub ACC (%)
Baseline	1.46M	89.96
SGC-GCN-A	2.06M	89.79
SGC-GCN-B	1.97M	89.78
SGC-GCN-C	1.52M	**90.39**

Within the framework of this experiment, we thoroughly examined the utilization of CTR-GCN as the baseline, and introduced SGC-GC as the fundamental building block to ensure a fair and equitable comparison. The outcomes of our experiments are meticulously presented in Table 1, where SGC-GC is gradually integrated into the baseline in three distinct categories: ascending order, non-ascending order, and global usage, based on the channels. The findings unequivocally demonstrate that the optimal performance of SGC-GCN is achieved when SGC-GC is selectively employed in the dimensionalization process, resulting in a remarkable increase in the accuracy rate. These outcomes unequivocally validate the efficacy of SGC-GC in enhancing the overall performance.

Table 2. The optimal model of the network structure is obtained by using one or multiple SGC-GCs or controlling the number of SGC-GCs with different activation functions, as shown in Table 2 in the experimental results section.

Methods	Param	P	X-sub ACC (%)
Baseline	1.46M	Relu	89.96
SGC-GCN-1	1.52M	Relu	**90.39**
SGC-GCN-2	1.65M	Relu	89.74
SGC-GCN-3	1.78M	Relu	89.94
SGC-GCN-4	1.90M	Relu	89.46
SGC-GCN-T	1.52M	Tanh	89.82
SGC-GCN-S	1.52M	**S**igmoid	89.69

In Table 2, we have separately compared the accuracy of SGC-GCN under different configurations. SGC-GCN-1 represents the use of 1 layer of SGC-GC, SGC-GCN-2 represents the use of 2 layers of SGC-GC, SGC-GCN-3 represents the use of 3 layers of SGC-GC, SGC-GCN-4 represents the use of 4 layers of SGC-GC. SGC-GCN-T represents the case where the activation function used is Tanh while using 1 layer of SGC-GC, and SGC-GCN-S represents the case where the activation function used is Sigmoid while using 1 layer of SGC-GC.

3.2. Comparison with the Latest Technologies

While numerous state-of-the-art methodologies incorporate a multistream fusion framework, we took an innovative approach by utilizing the CTR-GCN as our baseline model within the same framework, ensuring a fair and comprehensive comparison with other networks. Our approach combines four distinct models: articulated bone (AB), articulated motion (AM), and skeletal motion (SM). In Table 3, we present our NTU RGB+D 60 model, allowing for a direct comparison with the most recent techniques. Remarkably, our method surpasses these approaches in terms of X-sub evaluation, emphasizing its superior performance.

Table 3. A comprehensive analysis of classification accuracy was conducted, where our approach was compared against the most advanced techniques available, establishing a benchmark on the NTU RGB+D dataset.

Methods	NTU-RGB+D 60	
	X-sub	X-view
PA-LSTM[11]	62.90	70.30
C-CNN+MTLN[12]	79.60	84.80
ST-GCN[1]	81.50	88.30
AS-GCN[8]	83.30	93.30
UNIK[9]	86.26	93.01
2S-AGCN[2]	86.32	94.20
PA-ResGCN-B19[10]	86.63	94.33
MS-G3D[7]	87.29	94.32
CTR-GCN[6]	89.96	**95.08**
SGC-GCN(Ours)	**90.39**	95.01

4. Conclusion

In our research, we introduce an innovative approach called Channel Topological Graph Convolution (SGC-GCN) for enhancing skeleton-based action recognition. SGC-GCN integrates both dynamic and static topologies, addressing the issue of weaker features produced during the refinement process of dynamic shared topological graph convolution. By conducting rigorous experiments on the NTU RGB+D dataset, we substantiate the supremacy of SGC-GCN over the currently prevailing methods. Our results unequivocally highlight its exceptional performance in the domain of action recognition.

References

[1] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).

[2] Shi L, Zhang Y, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 12026-12035.

[3] Zhang P, Lan C, Zeng W, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]//proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1112-1121.

[4] Ye F, Pu S, Zhong Q, et al. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 55-63.

[5] Cheng K, Zhang Y, Cao C, et al. Decoupling gcn with dropgraph module for skeleton-based action recognition[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. Springer International Publishing, 2020: 536-553.

[6] Chen Y, Zhang Z, Yuan C, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13359-13368.

[7] Liu Z, Zhang H, Chen Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 143-152.

[8] Li M, Chen S, Chen X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3595-3603.

[9] Yang D, Wang Y, Dantcheva A, et al. Unik: A unified framework for real-world skeleton-based action recognition[J]. arXiv preprint arXiv:2107.08580, 2021.

[10] Song Y F, Zhang Z, Shan C, et al. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition[C]//proceedings of the 28th ACM international conference on multimedia. 2020: 1625-1633.
[11] Dai D, Xiao X, Lyu Y, et al. Joint extraction of entities and overlapping relations using position-attentive sequence labeling[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 6300-6308.
[12] Ke Q, Bennamoun M, An S, et al. A new representation of skeleton sequences for 3d action recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3288-3297.