MS-GTR: Multi-stream Graph Transformer for Skeleton-Based Action Recognition

Weichao Zhao[1,2], Jingliang Peng[1,2], and Na Lv[1,2(✉)]

[1] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, China
[2] School of Information Science and Engineering, University of Jinan, Jinan 250022, China
ise_lvn@ujn.edu.cn

Abstract. Skeleton-based action recognition has achieved remarkable progress by employing graph convolutional neural networks (GCNs) to model correlations among body joints. However, GCNs have limitations in establishing long-term dependencies and are constrained by the natural connections of human body joints. To overcome these issues, we propose a Graph relative TRansformer (GTR) that captures temporal features through learnable topology and invariant joint adjacency graphs. The GTR provides a high-level representation of the structure of the spatial skeleton, seamlessly integrated into the time series. Moreover, we introduce a Multi-Stream Graph Transformer (MS-GTR) to integrate various dynamic information for an end-to-end human action recognition task. The MS-GTR applies a double-branch structure, where the GTR is implemented as the main branch to extract long-term dynamic features, and an auxiliary branch processes short-term kinematic content. Finally, we use cross-attention as an inter-branch interaction mediator. Experimental results on the HDM05, NTU RGB+D, and NTU RGB+D 120 datasets demonstrate the potential of the proposed MS-GTR model for improving action recognition.

Keywords: Action recognition · Transformer · Skeleton-based · Graph

1 Introduction

Given its potential applications in video surveillance [13] and virtual reality [11], human action recognition has garnered significant interest from academia and industry. Compared to the original video data, skeleton data offers several advantages, including mitigating complex background interference and adapting to dynamic changes. Consequently, researchers have developed various skeleton-based action recognition methods. While existing action recognition methods exhibit diversity, there is a consensus that extracting sufficient spatial-temporal information is crucial. Traditional approaches commonly use handcrafted features to model the spatial human joint framework and dynamic information

B. Sheng et al. (Eds.): CGI 2023, LNCS 14497, pp. 104–118, 2024.
https://doi.org/10.1007/978-3-031-50075-6_9

in the temporal dimension. However, these exquisitely designed features are tailored to specific data and applications but are difficult to generalize. Deep learning techniques have rapidly evolved in recent years and are widely used for autonomous feature extraction. Representative networks include convolutional neural networks (CNNs) for processing static images and recurrent neural networks (RNNs) for modeling long-term contextual information in sequential data, such as joint coordinate sequences. In virtue of the peculiarity of the non-Euclidean data format of the natural physical connection of the skeleton structure, Yan et al. [32] pioneered graph-based approaches to model joints and their contacts for skeleton-based action recognition using graph convolutional neural networks (GCNs) and temporal convolution. Since then, GCNs have become the dominant deep neural network architecture for skeleton-based action recognition. Despite their success, GCNs still struggle to establish long-term temporal dependencies and often overlook joint cooperative relationships in motion. For instance, the "clapping" motion heavily relies on the cooperation between the left and right hands, but consciously focusing on the joint-to-joint relationship can lead to computational problems in the model.

In this paper, we proposed a novel framework for skeleton-based action recognition called Multi-Stream Graph Transformer (MS-GTR), as illustrated in Fig. 1. This framework enables effective multi-scale processing of skeleton information and extraction of representative spatio-temporal features. Concretely, we improve the transformer not only to model sequence context dependencies but also to incorporate the graph structure of the skeleton in action recognition. Additionally, we extract diverse information from the joint trajectories to enrich the range of expressions. We divide the data into the main and auxiliary branches to avoid computational complexity. While the main branch is always involved in feature extraction, various extra streams, *e.g.*, self-similarity matrices(SSM) and difference, provide dynamic short-range information to support the main feature extractor. We employ cross-attention for information exchange between the branches to facilitate the efficient integration of motion features across different scales.

As depicted in Fig. 1, the main unit provides only a token representing global information and interacts with the feature conveyed by the auxiliary branch using cross-attention. This token has absorbed the supplemental information, returns to the main unit, and undergoes subsequent operations. We conducted experiments on several human action datasets, including HDM05, NTU RGB+D, and NTU RGB+D 120, and the obtained results validate the value of our approach in improving action recognition performance.

Our main contributions to this work are summarized as follows:

1. A novel graph Transformer architecture is proposed to represent action sequences' higher-order spatial-temporal features and eliminate the redundant dependencies associated with fixed body connectivity.
2. We propose the multi-stream model called MS-GTR that consists of two distinct branches. The main branch is designed to extract the long-term dynamic features from the joint data directly. The auxiliary branch provides short-term information.

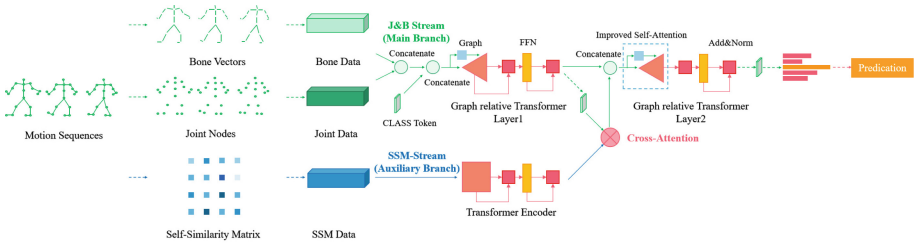

Fig. 1. Illustration of the proposed Multi-Stream Graph Transformer.

2 Related Work

2.1 Vision Transformer

The transformer [29] is a famous attention-based neural network architecture initially proposed for natural language processing. In addition to its success in NLP, the transformer has also proven its excellence for many fundamental computer vision tasks, *e.g.* classification [2,8,35], detection [1,12], and segmentation [30,36]. In particular, Zhang et al. [35] introduced the Video Transformer (VidTr) with spatio-temporal separable attention, which outperformed convolutional-based approaches for video classification. Sun et al. [27] built a multi-stream transformer network to model motion at different scales, taking advantage of the transformer's ability to capture long-range time dependencies. Chen et al. [2] applied a dual-branch vision transformer to complete the task of multi-scale feature extraction and image classification. Moreover, a simple and practical information exchange scheme between branches was proposed based on cross-attention. Inspired by their work, we offer a novel network to provide supplementary information for motion sequences at different scales through cross-attention. EAPT [47] proposes Deformable Attention, which learns offsets for each position in patches to obtain non-fixed attention information that can cover various visual elements.

2.2 Skeleton-Based Action Recognition

Skeleton-based action recognition aims to identify action through the human skeleton sequence. The most significant advantage of the first category of networks is that they take complete account of the long-term contextual associations in action. For example, Du et al. [9] fed the hierarchical structure of the human skeleton into an end-to-end hierarchical RNN, and the parts were reused and spliced together as the number of layers in the network increased. Two-stream temporal convolution networks proposed by Jia et al. [15] fully used inter-frame and intra-frame action characteristics. Xie et al. [18] proposed a temporal-then-spatial recalibration scheme that introduced the attention mechanism to recalibrate the temporal attention of frames and then further process using a convolutional neural network.

The above methods for skeleton-based action recognition primarily focused on capturing temporal features from human skeleton sequences. Still, they struggled to extract spatial characteristics from the topology of the connections between joints. Graph convolutional networks (GCNs) have emerged as a promising solution to this challenge. Yan et al. [32] were the first to apply GCNs to model dynamic skeletons for this task, but the graph topology heavily influenced the expressiveness of the model. Compared to manually setting fixed graph topology, Shi et al. [25] developed an adaptive GCN to learn the graph topology uniformly or individually. Cheng et al. [5] used parameterized topology for channel groups, but their model was bloated. Going a step further, Chen et al. [4] proposed a channel-wise graph convolution that shared a learnable topology as a generic prior for all channels and learned each channel-specific topology in a refinement way, which overcomes the inflexibility of previous methods like 2s-AGCN [25]. The adaptive graph convolutional block used in our proposed model to capture the spatial features is similar to the channel-wise methods. GAT [48] utilizes velocity information in a data-driven manner to learn discriminative spatial-temporal motion features from the sequence of skeleton graphs.

3 Method

3.1 Graph Relative Transformer

Motivation. Expressing higher-order spatial topology, adequately capturing contextual relationships, and effectively modeling spatial-temporal dependencies are essential for the signature representation of human action. However, unbiased modeling of long-term joint relationships can limit reliance on fixed natural connections in the human body, resulting in redundant dependency problems. In other words, due to the model's excessive focus on the genuine relationships of the body, the potential interactions between joints are easily overlooked. At the same time, the extraction of temporal features is over-reliant on the temporal convolution module. Adopting a fixed convolution kernel for feature extraction cannot adapt to feature changes in different periods, resulting in inadequate local feature extraction. As shown in Fig. 2, we aim to develop a graph topology that goes beyond the natural connectivity of the human body and can represent potential information of human pose. We use this topology to participate in the spatial-temporal feature with the improved Transformer. The goal is to capture long-term dependencies while retaining constraints on the higher-order spatial information of the skeleton in a lightweight manner.

Implementation of Graph. We attempt to find a reasonable and relatively accessible graph topology to guide us in constructing the spatial information of the skeleton. Our model involves two forms of graph convolution units to gather spatial details in a single frame. In the first approach, we follow the critical design of the spatial graph convolutional neural network proposed by [32]. The difference is that the sampling function is redefined using attention scores instead

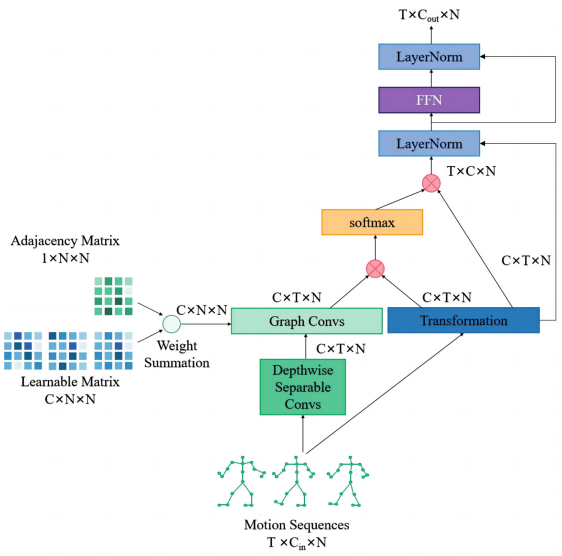

Fig. 2. Illustration of the proposed graph relative transformer.

of inter-joint connections, and the partition strategy is redesigned by manually setting thresholds to extract different scales of joint information. We determine the attention score among vertices as follows:

$$a_{ij} = \mathbf{x}_i \mathbf{w}_i \cdot (\mathbf{x}_j \mathbf{w}_j)^T \tag{1}$$

where $\mathbf{w}$ is the weight parameter. After normalization we end up with a $N \times N$ attention scores matrix and partition the vertex neighborhood by thresholds. This partition strategy can filter the vertices that are more related to a certain vertex as:

$$N(v_i) = \{v_j | threshold_{low} < a_{ij} \leq threshold_{up}\} \tag{2}$$

where $threshold_{up}$ and $threshold_{low}$ are the upper and lower limits of the thresholds, respectively. However, this approach dramatically increases the computational complexity of the model when calculating the attention scores, resulting in a waste of computational resources, and it does not show the expected effect during the experiment as shown in Sect. 4.3.

Inspired by [4], we aim to capture potential dependencies between joints in a manner that extends beyond the channels of the original spatial coordinates. Therefore, we designed a learnable matrix to represent the degree of association between joints, unlike the attention score matrix, which is parameterized and obtained through model optimization. However, as this solution may lead to the loss of some graph structure features by operating globally, we also add the adjacency matrix as a guide to natural connection to complement it. Another advantage of our approach is that the channels are grouped and globally averaged within the group pooling, which helps simplify the network model. As shown in

Fig. 2, relying on the joint association relationships provided by the learnable and adjacency matrices, we apply a graph convolutional operation on the features extracted by the depthwise separable convolution block to obtain an update of the nodal expression features. The obtained feature vectors will be directly involved in extracting the inter-frame dependencies.

Improved Transformer. The standard transformer is the backbone of our model, which we improve to achieve better performance and establish a baseline for subsequent experiments. To incorporate the spatial details of the skeleton, we reconstruct the query vector in the following way:

$$\mathbf{q}_i^t = \sum_{v_j^t \in N(v_i^t)} a_{ij}^t \phi_q(\mathbf{x}_j^t) \tag{3}$$

where a_{ij}^t denotes the weight coefficient of vertex i and vertex j at time t in the generalized sense. After fully considering the spatial feature map, we obtain each frame's corresponding key-value vector pairs by transforming the channel features using Eqs. (4) and (5):

$$\mathbf{k}_i^t = \phi_k(\mathbf{x}_i^t) \tag{4}$$

$$\mathbf{v}_i^t = \phi_v(\mathbf{x}_i^t) \tag{5}$$

In the temporal dimension, the focus is on modeling the frame-to-frame dependencies to obtain a representative feature map that incorporates the attention mechanism of the transformer:

$$\alpha_n^{ij} = \mathbf{q}_n^i \cdot {\mathbf{k}_n^j}^T \tag{6}$$

where i, j is the frame number of the action sequence, and n is the index number of the joint node. The joint characteristics of a given frame are updated based on information shared from the same joint in other frames, which establishes a strong long-term dependency. A global information token is introduced to summarize events for the entire time sequence, similar to the approach used in natural language processing. The update equation for a joint node's features in a given frame is as follows:

$$y_n^i = \sum \sigma(\frac{\alpha_n^{ij}}{\sqrt{d_k}}) v_n^j \tag{7}$$

where σ is an activation function that normalizes the input. After all the frames have been aggregated, the subsequent feed-forward layer can adjust the dimensions of the output, adding additional capabilities to the model.

3.2 Multi-stream Model Architecture

Based on GTR, we have constructed a robust MS-GTR model for integrating a variety of dynamic information streams toward skeleton-based action recognition. The overall model architecture is shown in Fig. 1. The proposed algorithm

operates on an action sequence $S = \{s^1, \ldots, s^t, \ldots, s^T\}$, consisting of T frames where each element $s^t \in \mathbb{R}^{N\times 3}$ represents the 3D coordinates of all available captured joints at a particular frame. We introduce the main and auxiliary branches in the model to capture a broader range of action details.

The main branch is concerned with the long-term dynamic information representation in the joint and bone data. Long-term dynamic information representation refers to a change in motion over long periods, usually in terms of modeling sequence contextual relationships. Specifically, we introduce bone representation as an interpretation of inter-joint connection to obtain directly from the original joint coordinates, which is then fed into the main branch along with the underlying joint features. To calculate bone vectors, which describe the relationship between two joints, we adopt the same approach as in [25]. Given a pair of head joint $J_i = \{x_i, y_i, z_i\}$ and tail joint $J_j = \{x_j, y_j, z_j\}$, we calculate the second-order information as $B_{i,j} = \{x_j - x_i, y_j - y_i, z_j - z_i\}$.

The auxiliary branch, in contrast, captures short-term features, such as the self-similarity matrix (SSM) and the difference between frames(also known as velocity). Given a set of joint features $J = \{J_1, J_2, \ldots, J_N\}$, we construct the self-similarity matrix $M_{ssm} \in \mathbb{R}^{N\times N}$ by comparing all the elements in the joint feature set with each other using the calculation formula $M(i, j) = SSM(J_i, J_j)$. The dot product of elements with each other is the simplest way to calculate the self-similarity matrix. This SSM data is used as the input of the auxiliary branch and as an additional information stream to the main branch regarding joint tightness. Likewise, the motion velocity of joints contains a wealth of action features. The velocity of a particular joint can be calculated as $\nu^t = s^{t+1} - s^t$, where ν^t is a vector reflecting the difference between two continuous frames in the original action sequence. This velocity information can be input into an auxiliary branch to support the main component regarding speed characteristics.

To ensure simplicity and fluency, we limit the role of the auxiliary branch to information transfer and rely on the main branch for capturing the most useful action features. To facilitate interaction between the main and auxiliary branches, we employ cross-attention. The remaining streams provide additional supplements to the main branch through the participation of a token related to global information. Specifically, the token of the main branch, $token_{main}$, is concatenated with the sequence data $S_{auxiliary} = \{s^1_{aux}, \ldots, s^T_{aux}\}$ arising from the auxiliary branch. Subsequently, a self-attention mechanism is implemented on the updated sequence $S_{fresh} = \{token_{main}, s^1_{aux}, \ldots, s^T_{aux}\}$, allowing the $token_{main}$ also to detect the characteristics of the auxiliary branch.

4 Experiments

4.1 Datasets

HDM05. HDM05 [22] is captured using optical marker-based technology, which helps to reduce noise interference in the motion capture data. It contains trajectories for 31 joints from 130 motion classes performed by five actors. And among

these 130 categories, some can be grouped into one category due to the same expression meaning, so we finally get the data of 65 action categories.

NTU RGB+D. The NTU RGB+D [23] involves the capture of motion sequences using three synchronized Microsoft Kinect v2 devices. The dataset contains 56,880 clips from 40 subjects, with each action organized into one of 60 action categories(including 11 multiplayer action categories). The skeleton data includes the 3D coordinates of 25 major joints at each frame. The dataset offers two evaluation criteria for action recognition methods: **Cross-View** is based on the camera's viewpoint that captured the action. The training set consists of 37,920 samples captured from a 45-degree view from the left and right, while the test set contains 18,960 samples captured from the front view. **Cross-Subject** validates the model in terms of different subjects. The experiment had 40 subjects categorized into training and test groups, each containing 20 actors. The training and test sets contain 40,320 and 16,560 samples, respectively.

NTU RGB+D 120. The NTU RGB+D 120 [40] is a large-scale dataset expended from the NTU RGB+D dataset. In addition to the 60 categories in the previous dataset, this dataset has an additional 60 types (i.e., 120 classes in NTU RGB+D 120). This dataset comprises 114,480 action clips captured from 155 camera views with 106 subjects. The authors of this dataset likewise recommend two benchmarks: Cross-Subject, similar to the previous dataset, are grouped by subjects, with 53 subjects in each group (63,026 samples for training and 50,922 clips for validation). **Cross-Set** is based on the setup made at the cameras' height and distance to the subjects to construct the training and testing set. The training set consists of 54,471 samples, while the test set contains 59,477 samples. The dataset includes 56,880 RGB+D video samples from 40 subjects, with each action classified into one of 60 action categories (including 11 multiplayer action categories).

4.2 Implementation Details

The implementation of our model is based on PyTorch and was run on an NVIDIA GeForce RTX 3090 GPU. We use gradient descent to update the model parameters. Specifically, we employed a stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.0004. We set the maximum number of training epochs to 120 and the size of each batch to 64. The initial learning rate is set to 0.01 and decays to 0.1 of the previous at epochs 45, 75, and 90. We employed the cross-entropy function as the loss function and added label smoothing to alleviate over-fitting and improve the model's generalization ability.

4.3 Ablation Study

We can demonstrate the contributions of the proposed components in GTR to achieve the goal of action recognition through a series of relevant experiments on the Cross-View benchmark of the NTU RGB+D dataset.

Graph Relative Transformer Block. First, we need to confirm whether it is necessary to use graph convolution to construct skeletal spatial guidelines instead of using transformers for the intuitive processing of time series. As previously mentioned, we feed the features extracted by the standard transformer to the classifier for action recognition and compare the result. Additionally, this section will discuss the implications of aggregating and updating graph nodes. As shown in Table 1, the experiment result indicates that combining the initial transformer model with the graph-related blocks significantly improves performance. Therefore, the most effective GTR associated with channels is selected as the foundational component for later model upgrades.

Table 1. Comparisons of the action recognition accuracy on Transformer with and without graph dependencies.

Methods	Accuracy (%)
Standard transformer	81.17
GTR (Threshold-dependent)	87.66
GTR (Learnable-matrix)	**89.74**

Table 2. Comparison of the effect of the presence or absence of auxiliary branch and different action input modalities on recognition results.

Joint	Bone	SSM	Difference	Accuracy (%)
√	×	×	×	90.33
×	√	×	×	88.06
√	√	×	×	90.32
×	×	√	×	90.68
√	×	√	×	91.19
×	×	×	√	85.17
√	√	√	×	**92.25**

Multi-stream Framework. To improve the representation and generalization ability of the model, we introduce an auxiliary branch to provide different forms of action implications to supplement the main unit. We conducted experiments by blocking the auxiliary streams and comparing their effects with those of the main branch alone. The results in Table 2 show that the fusion of information between branches can provide better semantic support for recognition. We also

measured the strength of the different action dynamic expressions introduced to the main branch for recognition. When only difference is available, the model performance is most unsatisfactory. Although the joint flow alone can improve the model to 90.33%, adding the self-similarity matrix to the main branch as an auxiliary information flow improves the recognition accuracy by 0.86%. It indicates that the collaboration of multi-stream information is more beneficial for the final action recognition.

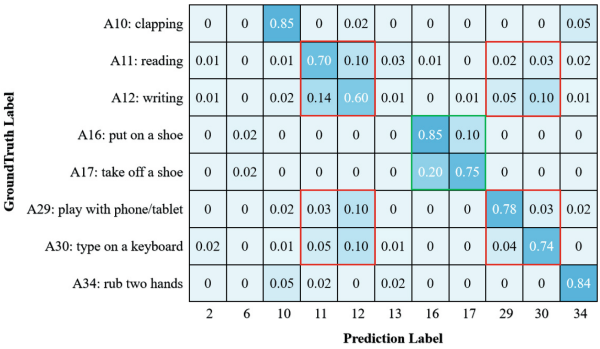

Fig. 3. Confusion matrix of skeleton-based action recognition with MS-GTR building on the cross View validation of NTU RGB+D dataset.

4.4 Confusion Matrix Analysis

As shown in Fig. 3, we visualized the confusion matrix of the cross-view benchmark results on the NTU RGB+D dataset to identify the categories that caused substantial interference leading to false recognition. Two situations can cause confusion between classification categories. The first set included categories where the inability to capture the reference led to some inaccuracy in recognition, including "A11:reading", "A12:writing", "A29:playing with phone or tablet", and "A30:type on a keyboard". These actions all involved manipulating the hands, and the specific tools used varied between the categories. The second set of confusing categories was the inverse order of each other, such as the pair of action

Table 3. Performance comparison on HDM05.

Methods	Accuracy (%)
Hierarchical RNNs (2015)	96.92
Deep LSTM (2016)	96.80
DHMR (2021) [39]	98.30
MANs (2021)	99.04
GTR (Ours)	**99.34**

sequences "A16:put on a shoe" and "A17:take off a shoe". We provided an explanation that as the network goes deeper, location information appears to become less significant.

Table 4. Performance comparison on NTU RGB+D dataset.

Methods	NTU RGB+D	
	Cross Subject (%)	Cross View (%)
Hierarchical RNNs (2015)	59.10	64.00
Clips + CNN + MTLN (2017) [16]	79.57	84.83
IndRNN (2018)	81.80	87.97
VA-RNN (2019) [33]	79.80	88.90
AMCGC-LSTM (2020) [31]	80.10	87.60
RGB+Skeleton (2020) [10]	84.23	89.27
TS-TCNs (2020)	82.40	90.20
MANs (2021)	79.74	91.55
ST-GCN (2018)	81.50	88.30
AM-STGCN (2019) [17]	83.40	91.40
2s-AGCN (2019)	88.50	95.10
Advanced CA-GCN (2020) [38]	83.5	91.4
LSGM+GTSC (2020) [14]	84.71	91.74
MS-G3D (2020) [41]	91.50	96.20
MST-GCN (2021) [42]	91.50	96.60
FV-GCN (2022) [28]	81.70	89.80
STAR (2021) [24]	83.40	89.00
MS-GTR (Ours)	84.50	92.25

Table 5. Performance comparison on NTU RGB+D 120 dataset.

Methods	NTU RGB+D 120	
	Cross Subject (%)	Cross Setup (%)
ST-LSTM (2016)	55.7	57.9
GCA-LSTM (2017)	61.2	63.3
FSNet (2019)	59.9	62.4
ST-GCN (2018)	70.7	73.2
2s-AGCN (2019)	82.9	84.9
MS-G3D (2020)	86.9	88.4
SGN (2020) [46]	77.9	78.5
MST-GCN (2021)	87.5	88.8
EfficientGCN (2022) [43]	88.3	89.1
PGT (2022) [45]	86.5	88.8
KA-AGTN (2022) [44]	86.1	88.0
MS-GTR (Ours)	78.3	80.8

4.5 Comparison to the State of the Art

To visually verify the feasibility and effectiveness of our model on action recognition, we conducted experiments on the HDM05 dataset (Table 3), the NTU RGB+D dataset (Table 4), and the NTU RGB+D 120 dataset (Table 5).

Notably, on the HDM05 dataset, we are currently at the forefront with a result of 99.34%. Whether it is the NTU RGB+D or the extended version, our model always has an advantage in recognition accuracy compared to the recurrent approaches, which indicates that our baseline model extracts superior features when establishing temporal dependencies. However, we still have a long way to go regarding a series of graph convolution variants of the method. Although our model is slightly less effective than 2 s-AGCN, we get a more significant improvement when we introduce the adaptive graph convolutional block, which proves the values of embedding the topology with the Transformer as the baseline model. For STAR [24], which was designed with the same intention as our baseline model, we used a graph structure to compensate for the lack of purely self-attention mechanisms to capture spatial features. Compared to this model, our recognition accuracy improved by 3.25% on Cross View and 0.75% on Cross Subject. In particular, taking KA-AGTN as an example, this model is also positioned as a graph transformer, but its model takes 2 s-AGCN as the baseline model and interpolates the attention layer for enhancing the dependence of local neighboring joints while preserving the spatio-temporal graph convolution layer. Our model starts from the most basic Transformer rather than using the currently available models as a baseline model to improve recognition accuracy, which results in us not fully utilizing the computational resources to represent the capabilities of our model. Therefore, to improve the model's generalization ability, we can use the existing pre-trained model to participate in the task, which is the direction we can improve in the future.

5 Conclusion

In this work, we propose a novel approach called GTR, which utilizes a transformer to efficiently capture the temporal features of action progression instead of solely relying on graph convolution neural networks. The proposed GTR involves a graph based on the natural connection of body parts in the expression update, which enhances the model expression diversification by motion features of different scales and makes the results more credible. We also introduce motion features with various expression meanings while reducing the complexity of model operations. In contrast to the direct fusion of action information from different scales, MS-GTR involves auxiliary input under the guidance of the main branch without introducing additional calculation costs. Our proposed MS-GTR achieves state-of-the-art performance on datasets captured by motion capture devices with widely varying accuracy and notably achieves leading recognition accuracy on the HDM05 dataset.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
2. Chen, C.F.R., Fan, Q., Panda, R.: CrossViT: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 357–366 (2021)
3. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13359–13368 (2021)
4. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13359–13368 (2021)
5. Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., Lu, H.: Decoupling GCN with DropGraph module for skeleton-based action recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12369, pp. 536–553. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_32
6. Cho, S., Maqbool, M., Liu, F., Foroosh, H.: Self-attention network for skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 635–644 (2020)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1110–1118 (2015)
10. Fan, Y., Weng, S., Zhang, Y., Shi, B., Zhang, Y.: Context-aware cross-attention for skeleton-based human action recognition. IEEE Access **8**, 15280–15290 (2020)
11. Fangbemi, A.S., Liu, B., Yu, N.H., Zhang, Y.: Efficient human action recognition interface for augmented and virtual reality applications based on binary descriptor. In: De Paolis, L.T., Bourdot, P. (eds.) AVR 2018. LNCS, vol. 10850, pp. 252–260. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95270-3_21
12. Gao, P., Zheng, M., Wang, X., Dai, J., Li, H.: Fast convergence of DETR with spatially modulated co-attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3621–3630 (2021)
13. Han, Y., Zhang, P., Zhuo, T., Huang, W., Zhang, Y.: Going deeper with two-stream ConvNets for action recognition in video surveillance. Pattern Recogn. Lett. **107**, 83–90 (2018)
14. Huang, J., Xiang, X., Gong, X., Zhang, B., et al.: Long-short graph memory network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 645–652 (2020)
15. Jia, J.G., Zhou, Y.F., Hao, X.W., Li, F., Desrosiers, C., Zhang, C.M.: Two-stream temporal convolutional networks for skeleton-based human action recognition. J. Comput. Sci. Technol. **35**(3), 538–550 (2020). https://doi.org/10.1007/s11390-020-0405-6

16. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3D action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3288–3297 (2017)
17. Kong, Y., Li, L., Zhang, K., Ni, Q., Han, J.: Attention module-based spatial-temporal graph convolutional networks for skeleton-based action recognition. J. Electron. Imaging **28**(4), 043032 (2019)
18. Li, C., Xie, C., Zhang, B., Han, J., Zhen, X., Chen, J.: Memory attention networks for skeleton-based action recognition. IEEE Trans. Neural Netw. Learn. Syst. **33**(9), 4800–4814 (2021)
19. 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)
20. Li, S., Li, W., Cook, C., Zhu, C., Gao, Y.: Independently recurrent neural network (IndRNN): building a longer and deeper RNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5457–5466 (2018)
21. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2684–2701 (2020)
22. Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., Weber, A.: Documentation mocap database HDM05 (2007)
23. 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)
24. Shi, F., et al.: STAR: sparse transformer-based action recognition. arXiv preprint arXiv:2107.07089 (2021)
25. 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 (CVPR) (2019)
26. Song, Y.F., Zhang, Z., Shan, C., Wang, L.: Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 1474–1488 (2022)
27. Sun, Y., Shen, Y., Ma, L.: MSST-RT: multi-stream spatial-temporal relative transformer for skeleton-based action recognition. Sensors **21**(16), 5339 (2021)
28. Tang, J., Wang, Y., Fu, S., Liu, B., Liu, W.: A graph convolutional neural network model with Fisher vector encoding and channel-wise spatial-temporal aggregation for skeleton-based action recognition. IET Image Proc. **16**(5), 1433–1443 (2022)
29. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
30. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
31. Xu, S., et al.: Attention-based multilevel co-occurrence graph convolutional LSTM for 3-D action recognition. IEEE Internet Things J. **8**(21), 15990–16001 (2020)
32. 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)
33. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive neural networks for high performance skeleton-based human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1963–1978 (2019)

34. Zhang, X., Xu, C., Tian, X., Tao, D.: Graph edge convolutional neural networks for skeleton-based action recognition. IEEE Trans. Neural Netw. Learn. Syst. **31**(8), 3047–3060 (2019)
35. Zhang, Y., et al.: VidTr: video transformer without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13577–13587 (2021)
36. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
37. Zhu, W., et al.: Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30 (2016)
38. Zhang, X, Xu, C, Tao, D.: Context aware graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14333–14342 (2020)
39. Lv, N., Wang, Y., Feng, Z., Peng, J.: Deep hashing for motion capture data retrieval. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2215–2219. IEEE (2021)
40. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L., Kot, A.: NTU RGB+ D 120: a large-scale benchmark for 3D human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2684–2701 (2019)
41. 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)
42. Chen, Z., Li, S., Yang, B., Li, Q., Liu, H.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1113–1122 (2021)
43. Song, Y., Zhang, Z., Shan, C., Wang, L.: Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 1474–1488 (2022)
44. Liu, Y., Zhang, H., Xu, D., He, K.: Graph transformer network with temporal kernel attention for skeleton-based action recognition. Knowl.-Based Syst. **240**, 108146 (2022)
45. Chen, S., Xu, K., Jiang, X., Sun, T.: Pyramid spatial-temporal graph transformer for skeleton-based action recognition. Appl. Sci. **12**(18), 9229 (2022)
46. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1112–1121 (2020)
47. Lin, X., Sun, S., Huang, W.: EAPT: efficient attention pyramid transformer for image processing. IEEE Trans. Multimedia **25**, 50–61 (2021)
48. Zhang, J., Xie, W., Wang, C.: Graph-aware transformer for skeleton-based action recognition. Vis. Comput. **39**, 4501–4512 (2023). https://doi.org/10.1007/s00371-022-02603-1