Spatio-temporal Attention Graph Convolutions for Skeleton-based Action Recognition

Cuong Le[1,2(✉)] and Xin Liu[1]

[1] Computer Vision and Pattern Recognition Laboratory, School of Engineering Science, Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland
cuong.le@liu.se, xin.liu@lut.fi

[2] Computer Vision Laboratory, Department of Electrical Engineering, Linköping University, Linköping, Sweden

Abstract. In skeleton-based action recognition, graph convolutional networks (GCN) have been applied to extract features based on the dynamic of the human body and the method has achieved excellent results recently. However, GCN-based techniques only focus on the spatial correlations between human joints and often overlook the temporal relationships. In an action sequence, the consecutive frames in a neighborhood contain similar poses and using only temporal convolutions for extracting local features limits the flow of useful information into the calculations. In many cases, the discriminative features can present in long-range time steps and it is important to also consider them in the calculations to create stronger representations. We propose an attentional graph convolutional network, which adapts self-attention mechanisms to respectively model the correlations between human joints and between every time steps for skeleton-based action recognition. On two common datasets, the NTU-RGB+D60 and the NTU-RGB+D120, the proposed method achieved competitive classification results compared to state-of-the-art methods. The project's GitHub page: STA-GCN.

Keywords: Computer vision · Action recognition · Graph convolution · Attention mechanism

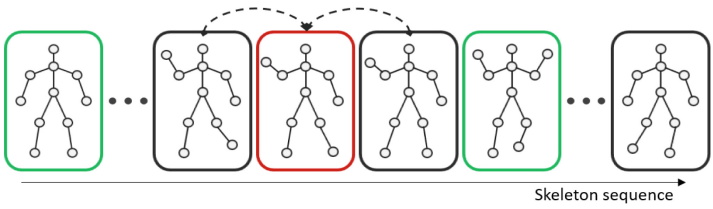
1 Introduction

Understanding human action plays a crucial role in many applications, such as video surveillance, human-computer interaction, and human behavior analysis [12,20,27,30]. Image-based methods suffer from many difficulties such as illumination changes, sensor noises, and perspective changes. One solution to overcome these problems in human action recognition is to utilize skeleton data.

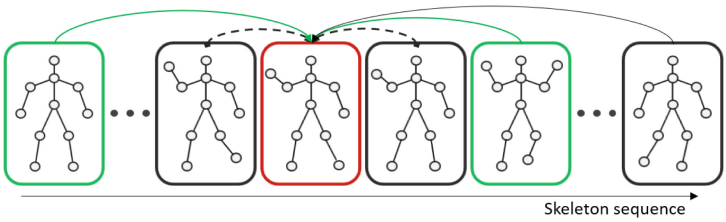
Earlier skeleton-based action recognition methods extract features using two main approaches: hand-crafted [8,13–15,25] and deep learning. Due to the recent development of large-scale datasets and computing power, deep learning approaches are providing better results in action recognition [27]. Earlier methods that use RNN [11,26] and CNN [5,6] overlook the correlation between joints

R. Gade et al. (Eds.): SCIA 2023, LNCS 13885, pp. 140–153, 2023.
https://doi.org/10.1007/978-3-031-31435-3_10

in the human skeleton, therefore limiting their expressive capability. Recently, graph convolutional networks (GCN) are introduced into skeleton modeling to integrate the natural human topology into the calculations [1–3,9,16,21,28,29], and they are currently state-of-the-art in skeleton-based action recognition. Besides, recent research often adapt the standard [9,21,28] or multi-scale dilated 1D convolution [1,2,16] to extract temporal features. However, in skeleton data, local consecutive frames usually contain similar features which are not useful to effectively discriminate challenging action samples. To overcome this, we could either increase the size of convolution filters or create a very deep model to widen the receptive field, but these methods can be computationally expensive. Another solution is to use the self-attention mechanism [23] that has the ability to pinpoint specific useful frames over a global receptive field, and this is the chosen approach of the paper. Figure 1 illustrates the proposed idea.



(a) Traditional temporal modeling with convolutions.



(b) Temporal modeling with added self-attentions.

Fig. 1. Temporal self-attention can pinpoint useful information along the action sequences by assigning weights on each frames. The red color denotes the current frame of calculation and the green frames indicate useful information. For many samples of human action, farther frames along the observed sequence can contain more discriminated features than the local neighborhoods. Aggregating local features is optimal for classifying challenging actions, so it is important to consider long-range dependencies. (Color figure online)

In this paper, a GCN-based attention method for skeleton-based action recognition is presented. In particular, a self-attention mechanism is combined with graph convolutional networks (GCN) for the spatial modeling of skeleton-based human models. In the temporal dimension, self-attention is applied together with multi-scale 1D convolutions to extract time-related features across the

skeleton sequences. The proposed model is tested on two large-scale datasets: NTU-RGB+D60 and NTU-RGB+D120, and the classification results proved to be competitive with the current state-of-the-art methods.

2 Related Works

2.1 Deep Learning on Graphs

Standard deep learning toolboxes are optimized for either grid-like data (CNN) or sequences (RNN), thus creating difficulties when applying them to graphs. Recently, GNN was introduced to define the learning task on graph-structure data [4]. Similar to traditional supervised learning methods, each skeleton sequence is represented as a data sample with an associated action label, and the goal is to learn the mapping from data points to labels. One of the GNN variants is graph convolutional networks (GCN) [7]. GCN is an approximation of spectral GNNs and is good at capturing graph features. However, the aggregation weights of GCN are explicitly defined based on node degrees, thus limiting its representation capability. Graph attentional networks (GAT) [24] is later introduced to address this problem by implicitly learning the connection weights through a self-attention mechanism. Both GCN and GAT are utilized in this study.

2.2 Skeleton-based Action Recognition

Earlier deep learning approaches for skeleton-based action recognition include RNN-based [11,26] and CNN-based [5,6] consider the skeleton graph as an uncorrelated set of features, and overlook the dynamic connectivity of the human body. GCN-based methods, that integrate the joint connections into their spatial modeling [1,2,9,16–18,21,22,28,29,31,32] record a significant boost in classification accuracy.

Yan et al. [28] first introduced the concept of GCN into action recognition, namely ST-GCN. The skeleton sequence is modeled from two types of edges: spatial edges that express connectivity between human joints, and temporal edges that connect the joints across time steps. Li et al. [9] proposed AS-GCN to capture richer dependencies in the spatial dimension of skeleton data. The method presents a module for capturing action-specific latent dependency between every human joint and extending human topology to represent higher-order dependencies. Shi et al. [21] reasoned that using predetermined and fixed skeleton graph topology for aggregating information is not optimal for diverse samples. Therefore, they proposed 2s-AGCN that captures second-order bone information in addition to joints' dependencies of skeleton data.

Liu et al. [16] proposed a disentangled and unifying graph convolutional network MS-G3D. The disentangling task removes the redundant dependencies between node features when aggregating spatial information and the graph topology is modified to directly obtain information from farther nodes. The combination of the two proposed methods creates a powerful extractor with

multi-scale receptive fields across spatial and temporal dimensions. Zhang et al. [31] also consider long-range dependencies by using context-aware graph convolutions (CA-GCN) based on self-attention. In addition to the local modeling of each joint vertex, CA-GCN integrates information from all other vertices within the sequence. Also relying on self-attention, Shi et al. [22] proposed decoupling schemes (DSTA-Net) to model spatio-temporal interactions between joints and frames without knowing their positions or mutual connections.

Chen et al. [1] proposed CTR-GCN that dynamically refine a shared prior topology for each feature channel. CTR-GCN creates multi-channel attention maps to refine the correlation between joints in each skeleton graph. This approach provides an effective modeling scheme from different channels, leading to stronger representation. Zhang et al. [32] propose a Spatial-Temporal Specialized Transformer Encoder (STST) to model the skeleton posture of each frame and capture changes of posture in the temporal dimension, thus providing strong modeling of action sequences. Similarly, various approaches [17,18] also utilizes the transformer architecture to extract spatial and temporal dependencies (ST-TR). Recently, Chi et al. [2] adopted the information bottleneck to derive the objective and the corresponding loss for maximum informative latent representation of skeleton-based actions.

In this paper, we explicitly combine the self-attention modeling of spatial and temporal dependencies from skeleton-based sequences. The proposed method consists of the strong spatial representation from the implicit interaction modeling between joints, and the ability to pinpoint useful time-based information of the temporal attention module.

3 Proposed Method

In this section, we first introduce the related notations on skeleton graphs and graph convolution. Detailed information about our proposed *Spatio-temporal attentional graph convolutions* is presented.

3.1 Preliminaries

A human action can be represented as a sequence of skeleton graphs. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an ordered pair constructed by a set of N vertices (or nodes) $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ and a set of edges $\mathcal{E}$ between these vertices. An edge going from node $u \in \mathcal{V}$ to node $v \in \mathcal{V}$ is noted as $(u, v) \in \mathcal{E}$. A graph can be conveniently formulated by an symmetry adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. The strength of edges is presented by the value of matrix's entries a_{ij}. The neighborhood of v_i is the set $\mathcal{N}(v_i) = \{v_j | a_{ij} \neq 0\}$. Action classes, represented as skeleton graph sequences, contain a node feature set $\mathcal{X}$, which can be presented in matrix form $\mathbf{X} \in \mathbb{R}^{T \times N \times C}$. The relationship between nodes within a frame is described by an adjacency matrix $\mathbf{A}$.

The calculation of GCN [7] adapts the idea of symmetric normalization into the node update function for an input skeleton feature $\mathbf{x}$ at layer k as:

$$\mathbf{h}^k = \sigma\big(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{x}^k\mathbf{W}\big) \tag{1}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with added self-loop, $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$, σ is the activation function, and $\mathbf{W}$ is the weight. Every entry of $\tilde{\mathbf{A}}$ takes the binary form to represent connectivity.

Graph Attentional Networks (GAT) [24] relaxes the entries of the adjacency matrix by adaptively learning it for every pair of connected nodes using the self-attention mechanism. The update function for an input skeleton feature $\mathbf{x}$ at layer k is formulated as:

$$\mathbf{h}^k = \sigma\left(\mathbf{M}\mathbf{x}^k\mathbf{W}\right) \tag{2}$$

where $\mathbf{M}$ is the self-attention score matrix which follows the calculation of [23]. In the original method, [24], the masked version of GAT is applied to only consider connected nodes. However, in this paper, the unmasked version of GAT is used instead to model the interactions between every pair of human joints.

3.2 Model Architecture

Our proposed feature extraction block consists of three modules connected to each other: GCN-GAT-combined spatial self-attention, temporal self-attention, and multi-scale temporal convolution. The proposed STA-GCN block is illustrated in Fig. 2. The architecture includes multiple blocks of STA-GCN, followed by a global average pooling, a fully connected, and a softmax layer.

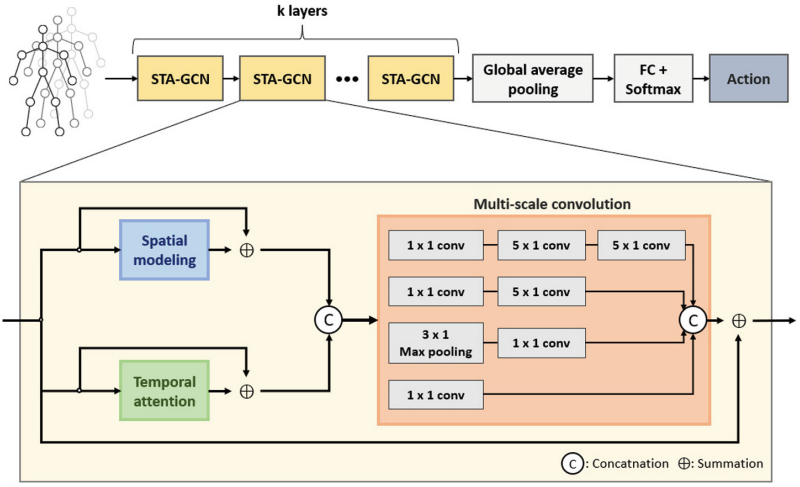


Fig. 2. Model architecture.

3.3 Spatial Modeling

To effectively extract features from the skeleton graphs, we combine the adjacency calculations from GCN and GAT into one update function:

$$\begin{aligned} \mathbf{C} &= \mathbf{M} + \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2} \\ \mathbf{h}^k &= \sigma\left(\mathbf{C}\mathbf{x}^k\mathbf{W}\right) \end{aligned} \tag{3}$$

where $\mathbf{C}$ is the combined adjacency matrix. By combining the two methods, we get the benefit from both. GCN is good for capturing spatial dependency between nodes from the given prior knowledge about human kinetics in the adjacency matrix. Unmasked GAT is good for modeling hidden correlations between human joints that are not visually connected.

The spatial modeling process is illustrated in Fig. 3. First, the input sequence is globally pooled along the temporal dimension. The pooled matrix is used to derive the query and key for computing the attention score. We then multiply the value with matrix $\mathbf{C}$ to get the final embedding tensor of the skeleton sequence. Furthermore, a multi-head version of self-attention is utilized to stabilize the $\mathbf{h}^k$ calculation. The combination of attention map and adjacency matrix also happens head-wised. Therefore, each head has its own combined attention map.

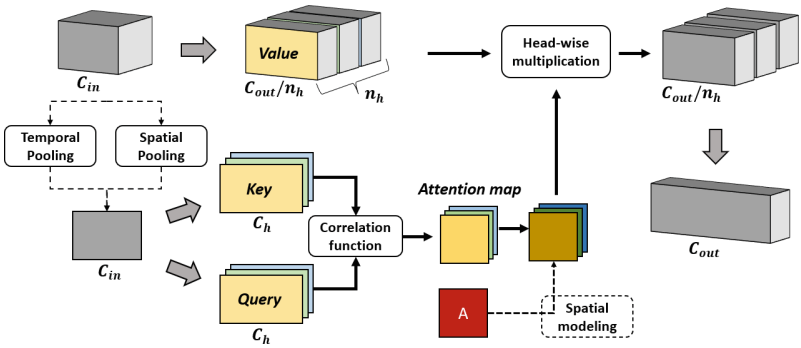


Fig. 3. Spatial and temporal self-attention modeling.

3.4 Temporal Modeling

The main goal of the temporal modeling process is to complement the traditional local extracting methods with a global aggregation approach to capture long-range dependencies within a skeleton sequence. Similar to the spatial modeling process, multi-head self-attention is also applied to temporally model skeleton sequences. First, each sequence is spatially pooled at every skeleton graph. After pooling, the data becomes a classical sequential problem. The attention map between frames is calculated from query and key. The only difference from the spatial modeling is no combination process with the adjacency matrix (Fig. 3). Then, the value is multiplied by the attention map to produce the output.

By using self-attention, we can extract long-range dependencies within one layer. The purpose of temporal self-attention is to pinpoint the most beneficial

frames from the skeleton sequence. However, the attention module may put a lot of weight into long-range frames and not consider local neighborhoods. As earlier methods demonstrated, extracting local features is an effective way to ensure a good baseline performance. While self-attentions pinpoint specific useful frames over the whole sequence, temporal convolutions provide direct access to neighborhood features. For this reason, we proposed combining self-attentions with 1D temporal convolutions to complement each other. We adopt the multi-scale dilated convolution module from [1,2,16] with minor changes to extract local temporal features. The dilated convolutions from [16] increase the receptive field while keeping the number of calculations unchanged. However, long-range dependencies are already collected by self-attentions, so standard 1D temporal convolutions are implemented instead to capture richer local dependencies.

4 Experiments

4.1 Datasets

In this study, we tested our algorithm in two datasets: NTU-RGB+D60 and NTU-RGB+D120.

NTU-RGB+D60 [19] is a large-scale dataset for action recognition that consists of 56,578 videos. The training samples are collected as skeleton sequences from 60 action classes, 40 distinct subjects, and 3 camera view angles. Four data modalities were provided but only 3D information of 25 body joints is used for the action recognition task. The authors propose two accuracy metrics: Cross-subject (Xsub) and Cross-view (Xview). In Xsub, 40 subjects are split evenly into training and testing sets. In Xview, samples from cameras 2 and 3 are used for training and camera 1 for testing.

NTU-RGB+D120 [10] is the extension of previous dataset. The updated version provides the addition of 57,367 skeleton sequences over 60 extended action classes. In total, NTU-RGB+D 120 consists of 113,945 training samples over 120 action classes, which are performed by 106 human subjects and captured from 32 different camera setups. The authors propose two evaluation settings: Cross-subject (Xsub) and Cross-setup (Xset). In Xsub, 106 subjects are split evenly into training and testing sets. In Xset, 32 collection setups are split as even-IDs for training and odd-IDs for testing.

4.2 Implementation Details

The experiment results are conducted on two NVIDIA Tesla V100 GPUs from Finland's CSC server with PyTorch deep learning framework. The model is trained with SGD with momentum 0.9, weight decay 0.0004, batch size 64, and an initial learning rate of 0.1 for 65 epochs. The learning rate is scheduled to decay with a rate of 0.1 at epochs 35 and 55. A warm-up strategy is adopted for the first 5 epochs to stabilize the training process. For two datasets NTU-RGB+D 60 and NTU-RGB+D 120, the data pre-processing from [1] is used, and all skeleton sequences are resized to 64 frames each.

Similar to [1,21,29], multiple training with different data modalities are also implemented. In addition to the original data of skeleton joints, bone and velocity modalities are also utilized for training. Thus, there are a total of four different training at each evaluation setting: joint, bone, joint-motion, and bone-motion. The performances of all four modalities are then assembled into a final value of accuracy.

The standard evaluation metric on human action recognition research is accuracy measurement. For a fair comparison, we also measured the classification accuracy on both NTU-RGB+D60 and NTU-RGB+D120. In addition, we also conduct F1 measurement over classes on some related methods that published their model's weights [1,16].

To find the best performing model for STA-GCN, an ablation experiment was conducted. With the same training hyper-parameters, a baseline model consisting of only graph convolutional networks and temporal convolutions was tested. Then, the proposed spatial adaptive attention and temporal attention were added respectively. All ablation study experiments were conducted in the cross-subject setting of the NTU-RGB+D60 dataset.

4.3 Ablation Study

In this section, the proposed spatial-temporal attention graph convolution network is tested on the cross-subject evaluation setting on the NTU-RGB+D60 dataset. An architecture similar to ST-GCN [28] is deployed as the starting baseline. There are three modules that need to be studied: adaptive GCN, additional GAT, and temporal attention modeling. Also, the original 1D convolution in ST-GCN [28] is changed to the multi-scale module for a fair comparison. The experimental results are shown in Table 1.

Table 1. Accuracy comparison for ablation study.

Methods	**Params.**	**Accuracy** (%)	**Mean last 10 epochs** (%)
Baseline	843868	87.50	87.27
Baseline w. adaptive GCN	850118	89.30	89.09
STA-GCN w/o. temp attention	1119422	89.88	89.72
STA-GCN w. temp attention	1174560	89.90	89.87

First, we tested the performance of the original baseline model with normal GCN and multi-scale convolutions. Then, the adaptive characteristic is integrated into the GCN module to observe the improvement. In the third experiment, GAT and self-attentions are fused into the spatial modeling to create the proposed STA-GCN model. Self-attentions along temporal dependencies are separately considered in the fourth experiment. It can be observed that the accuracy of the classifiers increases gradually as more modules are added.

With an overall accuracy of 87.50%, the baseline performs fairly well on the NTU-RGB+D60 dataset. However, the baseline model struggles against many difficult classes, such as: eating snack (action 02), reading (action 11), writing (action 12), taking off shoes (action 17), playing with phone (action 29), sneezing (action 41). These classes differ by small changes in arm movements, thus creating challenges for skeleton-based action recognition. By integrating the proposed modules into the baseline model, an increase in performance is recorded, as shown in Fig. 4.

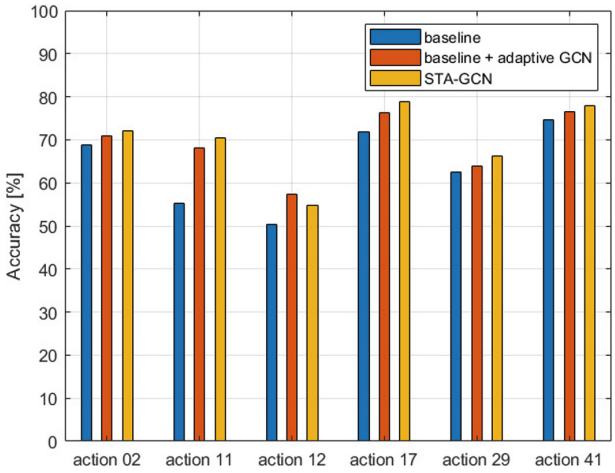


Fig. 4. Comparison of performance on difficult classes.

In addition, we measured the class-wise F1-score for the baseline and STA-GCN. Table 2 illustrates the top five classes with the highest improvement on F1 measurement. The highest improvements was recorded on classes that include small-gesture action samples such as reading (action 11), using a hand-fan (action 49), touching neck (action 47), or putting on glass (action 18). The action reading (action 11) has the highest improvement when STA-GCN was applied with 8% increment in F1-score. This demonstrates the positive impact when applying temporal attention modules to recognize subtle gesture action samples.

Figure 5 shows the last 10 epochs of the training process. In the case of the model without temporal attention, the accuracy oscillates between 89.6% to 89.9%, with a mean accuracy of 89.72%. While the performance of the model with temporal attention is stable at around 89.9% and an average of 89.87%. Therefore, though the final accuracy of the two models is the same, the one with temporal attention proved to be more consistent and superior. Because of these reasons, the STA-GCN module with temporal attention is chosen as the main building block of the final skeleton-based action recognition model.

Table 2. Top five classes with highest F1-score improvement.

	STA-GCN vs baseline		STA-GCN vs adaptive baseline	
Rank	**Action**	**Increment(%)**	**Ation**	**Increment(%)**
Top 1	action 11	8.05	action 11	5.13
Top 2	action 49	7.47	action 47	3.20
Top 3	action 47	7.12	action 18	2.98
Top 4	action 18	5.73	action 05	2.75
Top 5	action 29	5.66	action 28	2.71

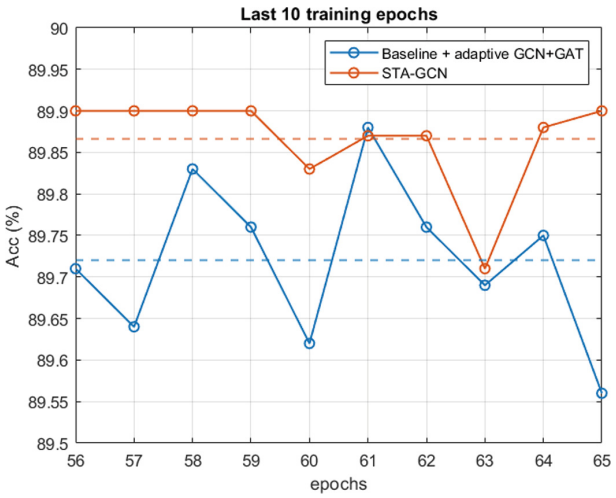


Fig. 5. Comparison between models with and without temporal attention.

4.4 Results

We adopt the same multi-stream fusion framework as [1,21], by fusing four different modalities of data: joint, bone, joint motion, and bone motion. The comparisons in classification accuracy of our approach with other graph-based methods is demonstrated in Table 3.

On NTU-RGB+D60, the final STA-GCN model beats earlier methods [6,11,25], and some recent GCN-based methods such as [9,16,21,28,29], but cannot outperform the current state-of-the-arts [1,2] in Xsub and Xview. However, compared to other methods that also implemented self-attention on temporal domain [17,18,22,31,32], our STA-GCN has a clear advantage, especially on Xsub setting. For NTU-RGB+D 120 dataset, the evaluation quality is the same. STA-GCN still outperforms [17,18,22,31,32] in both Xsub and Xset metrics. When compared to ST-TR [17], the method that is most similar to our approach, STA-GCN achieved closed performance on the NTU-RGB+D60 but outperforms by a large margin on the NTU-RGB+D120. Besides, STA-GCN

under-performs CTR-GCN [1] by 0.4% and 0.2% and InfoGCN by 1.3% and 0.8% on Xsub and Xset settings of the NTU-RGB+D120.

Table 3. Classification accuracy compared with state-of-the-art methods on the NTU-RGB+D60 and NTU-RGB+D120 datasets.

Methods	**Year**	**NTU-RGB+D60**		**NTU-RGB+D120**	
		Xsub (%)	**Xview** (%)	**Xsub** (%)	**Xset** (%)
Lie Group [25]	2014	50.1	52.8	–	–
Temporal CNN [6]	2017	74.3	83.1	–	–
Ind-RNN [11]	2018	81.8	88.0	–	–
ST-GCN [28]	2018	81.5	88.3	–	–
AS-GCN [9]	2019	86.8	94.2	–	–
2s-AGCN [21]	2019	88.5	95.1	82.9	84.9
MS-G3D [16]	2020	91.5	96.2	86.9	88.4
CA-GCN [31]	2020	83.5	91.4	–	–
DSTA-Net [22]	2020	91.5	96.4	86.6	89.0
Dynamic GCN [29]	2020	91.5	96.0	87.3	88.6
STST [32]	2021	91.9	96.8	–	–
ST-TR [17]	2021	89.9	96.1	81.9	84.1
CTR-GCN [1]	2021	92.4	96.8	88.9	90.6
Qin's method [18]	2022	90.5	96.1	85.7	86.8
InfoGCN [2]	2022	93.0	97.1	89.8	91.2
STA-GCN (Net)		92.4	96.5	88.5	90.4
STA-GCN (Joint+Bone)		92.0	96.4	88.4	90.0
STA-GCN (Joint)		89.9	94.9	84.7	86.3
STA-GCN (Joint-motion)		87.4	93.3	81.4	83.1
STA-GCN (Bone)		90.3	94.9	86.3	87.8
STA-GCN (Bone-motion)		87.4	91.9	81.2	83.0

As previously mentioned, F1-score measurement is also carried out for some related methods that published their model weights, specifically MS-G3D [16] and CTR-GCN [1]. The measurement is shown in Table 4. Because NTU-RGB+D60 and NTU-GRB+D120 is fairly balanced datasets, the F1-scores did not vary much from the accuracy measurements.

Table 4. F1-score measurement results

Methods	**NTU-RGB+D60**		**NTU-RGB+D120**	
	Xsub-joint (%)	**Xsub-bone** (%)	**Xsub-joint** (%)	**Xsub-bone** (%)
MS-G3D [16]	89.4	90.1	83.8	86.0
CTR-GCN [1]	89.9	90.5	85.2	85.7
STA-GCN	89.9	90.2	84.9	86.4

5 Conclusion

In this study, a spatio-temporal attentional graph convolution network for skeleton-based action recognition is presented. As the results attest, the combination of GCN, self-attentions, and temporal convolutions can effectively learn features and joint relationships from the action sequences. Compared to the state-of-the-arts on two common datasets NTU-RGB+D60 and NTU-RGB+D120, the proposed model STA-GCN achieved competitive classification performance. In the future, we believe it is worthwhile to put more focus on the modeling of micro-movements in human limbs, in order to increase the classification performance of the model in those challenging situations.

References

1. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13359–13368 (2021)
2. Chi, H.G., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: INFOGCN: representation learning for human skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20186–20196 (2022)
3. Gao, R., Liu, X., Yang, J., Yue, H.: CDCLR: llip-driven contrastive learning for skeleton-based action recognition. In: 2022 IEEE International Conference on Visual Communications and Image Processing (VCIP), pp. 1–5. IEEE (2022)
4. Hamilton, W.L.: Graph representation learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **14**(3), 1–159 (2020)
5. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4570–4579 (2017)
6. Kim, T.S., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1623–1631 (2017)
7. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR) (2017)
8. Koniusz, P., Cherian, A., Porikli, F.: Tensor representations via kernel linearization for action recognition from 3D skeletons. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 37–53. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_3

9. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3590–3598 (2019)
10. 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. (PAMI) **42**(10), 2684–2701 (2020)
11. Liu, J., Shahroudy, A., Xu, D., Kot, A.C., Wang, G.: Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **40**(12), 3007–3021 (2018)
12. Liu, X., Shi, H., Chen, H., Yu, Z., Li, X., Zhao, G.: imigue: an identity-free video dataset for micro-gesture understanding and emotion analysis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10631–10642 (2021)
13. Liu, X., Shi, H., Hong, X., Chen, H., Tao, D., Zhao, G.: Hidden states exploration for 3d skeleton-based gesture recognition. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1846–1855 (2019)
14. Liu, X., Shi, H., Hong, X., Chen, H., Tao, D., Zhao, G.: 3d skeletal gesture recognition via hidden states exploration. IEEE Trans. Image Process. **29**, 4583–4597 (2020)
15. Liu, X., Zhao, G.: 3d skeletal gesture recognition via discriminative coding on time-warping invariant riemannian trajectories. IEEE Trans. Multimedia **23**, 1841–1854 (2021)
16. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 140–149 (2020)
17. Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, Part III, pp. 694–701 (2021)
18. Qin, X., Cai, R., Yu, J., He, C., Zhang, X.: An efficient self-attention network for skeleton-based action recognition. Sci. Rep. **12**, 2045–2322 (2022)
19. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: a large scale dataset for 3d human activity analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1010–1019 (2016)
20. Shi, H., Peng, W., Chen, H., Liu, X., Zhao, G.: Multiscale 3d-shift graph convolution network for emotion recognition from human actions. IEEE Intell. Syst. **37**(4), 103–110 (2022)
21. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12018–12027 (2019)
22. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In: Asian Conference on Computer Vision (ACCV), pp. 38–53 (2020)
23. Vaswani, A., et al.: Attention is all you need. In: International Conference on Neural Information Processing Systems (NIPS), pp. 6000–6010 (2017)
24. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representations (ICLR) (2018)
25. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 588–595 (2014)

26. Wang, H., Wang, L.: Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3633–3642 (2017)
27. Wang, L., Huynh, D.Q., Koniusz, P.: A comparative review of recent kinect-based action recognition algorithms. IEEE Trans. Image Process. **29**, 15–28 (2020)
28. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI Conference on Artificial Intelligence, pp. 3482–3489 (2018)
29. Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., Tang, H.: Dynamic GCN: context-enriched topology learning for skeleton-based action recognition. In: ACM International Conference on Multimedia, pp. 55–63 (2020)
30. Yu, Z., et al.: Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. IEEE Trans. Image Process. **30**, 5626–5640 (2021)
31. Zhang, X., Xu, C., Tao, D.: Context aware graph convolution for skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14321–14330 (2020)
32. Zhang, Y., Wu, B., Li, W., Duan, L., Gan, C.: STST: Spatial-temporal specialized transformer for skeleton-based action recognition. In: ACM International Conference on Multimedia, pp. 3229–3237 (2021)