Spatiotemporal Graph Autoencoder Network for Skeleton-Based Human Action Recognition

Hosam Abduljalil *, Ahmed Elhayek, Abdullah Marish Ali, Fawaz Alsolami

Posted Date: 29 January 2024

doi: 10.20944/preprints202401.1998.v1

Keywords: graph convolutional networks; graph autoencoder; deep learning; human activity analysis; skeleton-based action recognition

Article

Spatiotemporal Graph Autoencoder Network for Skeleton-Based Human Action Recognition

Hosam Abduljalil [1,*], Ahmed Elhayek [2], Abdullah Marish Ali [1] and Fawaz Alsolami [1]

[1] Department of Computer Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2] Department of Artificial intelligence, University of Prince Mugrin, Medina 40202, Saudi Arabia
* Correspondence: haliabduualil@stu.kau.edu.sa (H.A.)

Featured Application: The recognition and interpretation of human actions play a crucial role in various practical applications such as video surveillance footage, healthcare systems, robotics field, human-computer interaction, etc.

Abstract: Skeleton-based human action recognition is a challenging yet important technique because of its wide range of applications in many fields, including patient monitoring, security surveillance, and observing human-machine interactions. Many algorithms that attempt to distinguish between many types of activities have been proposed. However, most practical applications require highly accurate detection of specific types of activities. In this study, a novel and highly accurate spatiotemporal graph autoencoder network for skeleton-based human action recognition is proposed. Furthermore, an extensive study was conducted using different modalities. For this purpose, a spatiotemporal graph autoencoder that automatically learns spatial as well as temporal patterns from human skeleton datasets was built. The powerful graph convolutional network named GA-GCN, developed in this study, notably outperforms most of the existing state-of-the-art methods based on two common datasets, namely NTU RGB+D [1] and NTU RGB+D 120 [2]. On the first dataset, we achieved an accuracy of 92.3% and 96.7% based on the cross-subject and cross-view evaluations, respectively. On the other more challenging dataset (i.e. NTU RGB+D 120), GA-GCN achieved 88.8% and 90.4% based on the cross-subject and cross-set evaluation, respectively.

Keywords: Graph convolutional networks, graph autoencoder, deep learning, human activity analysis, skeleton-based action recognition

1. Introduction

Recognition and analysis of human actions is a critical subfield in computer vision and deep learning, with the primary goal of automatically detecting and classifying human actions or gestures from video data. Sophisticated algorithms and models that can understand and interpret the dynamics of human movements are required for this purpose. The recognition and interpretation of human actions play a crucial role in various practical applications such as video surveillance footage, healthcare systems, robotics field, human-computer interaction, etc. Extracting meaningful information from video sequences enables machines to understand and respond to human actions, thereby enhancing the efficiency and safety of many domains.

This emerging field leverages the capabilities of deep learning techniques, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and graph convolutional neural networks (GCNs), to capture temporal and spatial features in video data. Recent advancements in the use of 3D graph convolutional neural networks (3D GCNs) have further improved the accuracy of action recognition models. Notable datasets such as NTU RGB+D [1], NTU RGB+D 120 [2], NW-UCLA, and Kinetics [3], have become benchmarks for evaluating and analyzing the performance of these techniques, driving further research and innovation in the field.

 doi:10.20944/preprints202401.1998.v1

1.1. Introduction to Skeleton-based Human Motion Prediction

The aim of using action prediction algorithms is to anticipate the classification label of a continuous action based on a partial observation along the temporal axis. Predicting human activities prior to their full execution is regarded as a subfield within the wider scientific area of human activity analysis. This field has garnered significant scholarly interest owing to its extensive array of applications in the domains of security surveillance, observing human-machine interactions, and medical monitoring. [4].

According to biological researches [5], human skeleton data as depicted in Figure 1, are sufficiently informative to represent human behavior, despite the absence of appearance information. Human activities are inherently conducted within a three-dimensional spatial context, making three-dimensional skeletal data an appropriate means of capturing human activity. The acquisition of 3D skeletal information can be efficiently and conveniently achieved in real-time through the utilization of affordable depth sensors, such as Microsoft Kinect and Asus Xtion. Consequently, the utilization of 3D skeleton data for activity analysis has become known prominent area of scholarly studies [6–8]. Among the advantages of this type of activity analysis are its conciseness, sophisticated representation, and resilience to differences in views, illumination, and surrounding visual distractions.

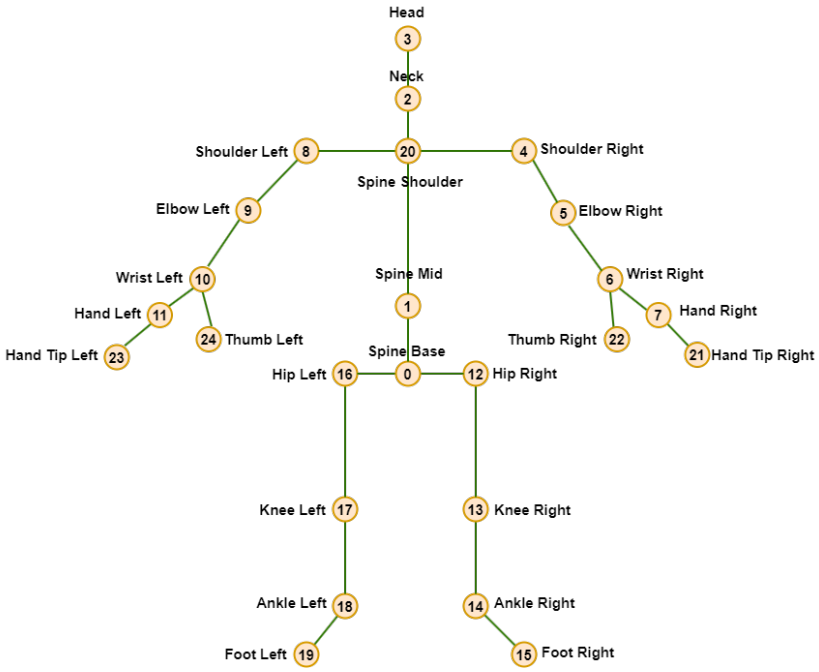

Figure 1. Human skeleton with circles representing joints and lines representing bones.

1.2. Description of the Work

In this study, the focus is only on human action recognition. Thus, the goal was to detect the motion of a person. Surveillance cameras can be found in almost all sensitive locations. However, the resulting number of video streams is not only difficult to monitor but also costly to transmit and store.

We performed the experiments using a well-known action recognition dataset named NTU RGB+D 120 [2], which extends NTU RGB+D [1] and provides the extracted skeletal motion of 120 motion classes.

The objective of this study was to develop a novel and highly accurate human motion detection algorithm. For this purpose, we focused on special practical scenarios and utilized the latest deep learning (DL) technologies (e.g. graph convolutional networks, autoencoders, and one-class classifiers) to develop a highly accurate human action recognition algorithm (for example, abnormal behavior of customers in shopping places can be detected by analyzing their motion based on 3D skeletal data).

 doi:10.20944/preprints202401.1998.v1

1.3. Graph Autoencoders

Graph autoencoders (GAEs) are a class of neural network models designed for learning low-dimensional representations of graph-structured data. In recent years, they have gained a notable focus due to their ability to perform tasks in a range of applications, including node classification, link anticipation, and community discovery. GAEs use the power of autoencoders to encode graph nodes into lower-dimensional latent representations and decode them back to the original graph structure. This process involves capturing both the topological structure and node attributes, making them powerful tools for graph representation learning. Notable works in this field include the GraphSAGE model by Hamilton et al. [9] and the variational graph autoencoder (VGAE) proposed by Kipf and Welling [10]. These models provide valuable insights into the development of graph autoencoders for various graph-related tasks.

In addition to these studies, there is an expanding body of research exploring variations and applications of graph autoencoders. These have been adapted for semi-supervised learning, recommendation systems, and anomaly detection. The field of graph autoencoders continues to evolve, offering promising avenues for further research and development.

1.4. The Basic Description of the Graph Autoencoder Skeleton-based Human Action Recognition Algorithm

GAEs refer to a class of neural network models that may be trained in an end-to-end manner. These models are specifically designed for unsupervised learning jobs, such as clustering and link prediction, on graph-structured data. GAEs are based on GCNs. We adopted the architecture of CTR-GCN that used a refinement way to learns channel-wise topologies which proposed by Chen et al. [11] as the base unit and trained a graph auto-encoder to automatically learn spatial as well as temporal patterns from data. Figure 2 illustrates the proposed network architecture. The input to the network is a spatial temporal graph based on skeletal sequences which can be generated as described by Cai et al. [12]. This graph is fed to the autoencoder for reconstruction. Finally, the verification process involves the application of thresholding to the reconstruction loss [13].

The main contributions of this study are the following:

- A novel spatiotemporal graph-autoencoder network for skeleton-based human action recognition; our GA-GCN pipeline is illustrated in Figure 2. Additional skip connections were incorporated to improve the learning process by enabling the direct flow of information from earlier layers to later layers.
- Outperforming most of the existing methods on two common skeleton-based action recognition datasets.
- Achieving notable improvement in the performance by introducing additional multiple modalities; see the experimental section 4.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 January 2024 doi:10.20944/preprints202401.1998.v1

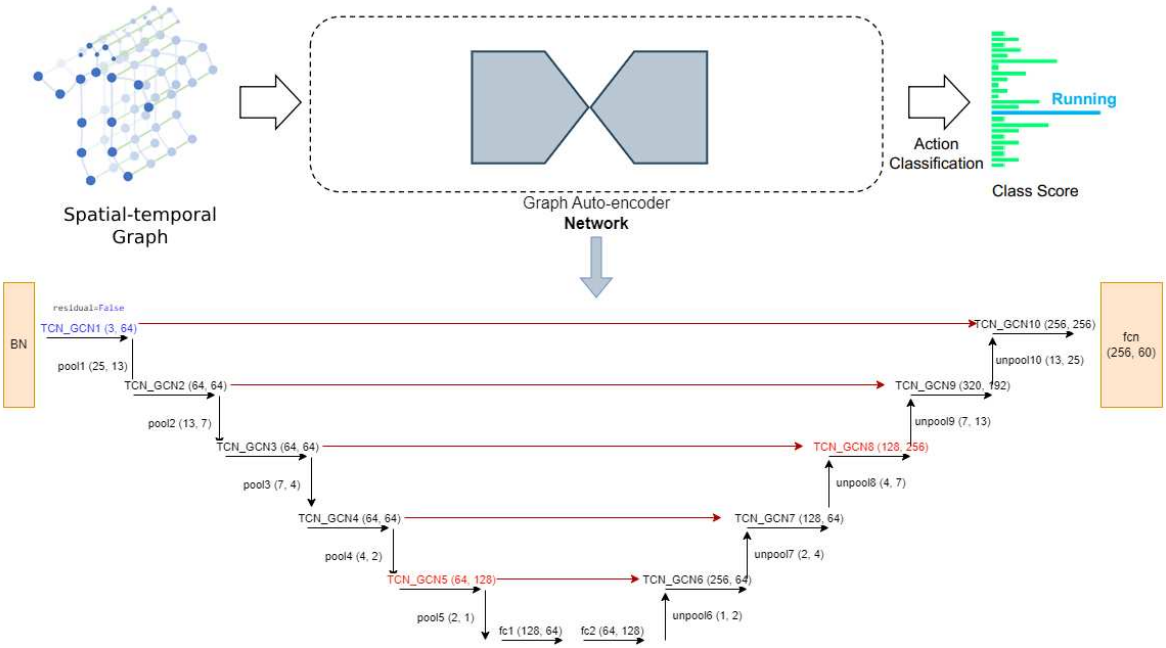

Figure 2. Network architecture schematic overview of the proposed spatiotemporal graph autoencoder network for skeleton-based human action recognition. The top figure is an overview of the entire pipeline and the lower figure shows the input channels and joints of the graph autoencoder parts and provides details of the pipeline layers. The input to this pipeline is a spatiotemporal graph which combines multiple poses of the sample video. This spatiotemporal graph is fed into the graph autoencoder network to produce the final output which is the probability of each class. The layers in the encoder part are skip connected and concatenated with layers in the decoder part (these are indicated by red lines in the above diagram).

2. Related Work

2.1. Graph Convolutional Networks

CNNs have produced impressive results in processing images. There is growing interest in creating GCNs. Spectral techniques perform convolution within the spectral domain [14–16]. Their application relies on Laplacian eigenbasis. Consequently, these methods are primarily suitable for graphs that share consistent structural configurations.

Convolutions are defined by spatial methods directly on a graph [17–19]. Handling various neighborhood sizes is one of the difficulties associated with spatial approaches. GCN proposed by Kipf et al. [16] is one of many GCN versions available, and it is widely adapted for diverse purposes because of its simplicity. Feature update rule of GCN comprises of transformation of features into abstract representations step and feature aggregation step based on the analysis of the graph topology. The same rules were used in this study for feature updates.

2.2. GCN-based Skeleton Action Recognition

The feature update rule proposed by Kipf et al. [16], which is often followed by GCNs, has been successfully adapted for recognition of skeleton-based action [20–26]. Numerous GCN-based techniques place a strong emphasis on topology modeling due to the significance of topology in GCNs. Based on topological variations, GCN-based techniques can be categorized into the following categories:

 doi:10.20944/preprints202401.1998.v1

1. Static and dynamic techniques: In static techniques, the topologies of GCNs remain constant throughout the inference process, whereas they are dynamically inferred throughout the inference process for dynamic techniques.
2. topology shared and topology non-shared techniques: Topologies are shared across all channels in topology shared techniques, whereas various topologies are employed in various channels or channel groups in topology non-shared techniques.

In the context of static approaches, Yan et al. [24] proposed a ST-GCN that predefined topology that is fixed during the training and testing steps in accordance with the human body structure. Multi-scale graph topologies were incorporated into GCNs by Liu et al. [20] and Huang et al. [27] to facilitate the modeling of joint relationships across many ranges.

In the context of dynamic approaches, Li et al. [6] suggested an A=links inference component to record correlations related to actions. Enhancing topology learning was demonstrated by Shi et al. [21] and Zhang et al. [28] by using a self-attention method that emulates the association between two joints. These techniques use regional information to infer the relationships between two joints. A dynamic GCN method for learning correlations between joint pairs was proposed by Ye et al. [25] by incorporating contextual data from joints. Dynamic methods offer greater generalization capabilities than static methods because of their dynamic topologies.

Methods with and without a common topology. Static and dynamic topologies are shared across all channels in topology sharing procedures. These strategies impose limitations on the performance of models by compelling GCNs to aggregate features across channels that possess identical topology. The majority of GCN-based techniques, including the aforementioned static methods [20,24,27] and dynamic methods [6,21,25,28], operate in a topology shared manner.

Topology non-shared techniques naturally overcome the drawbacks of topology shared techniques by employing various topologies in various channels or channel groups. A decoupling graph convolutional network (DC-GCN) was proposed by Cheng et al. [29] that establishes unique parameterized topologies for various channel groups. DC-GCN encounters optimization challenges when constructing channel-wise topologies caused by too many parameters. In skeleton-based action recognition, topology non-shared graph convolutions have rarely been investigated. Chen et al. [11] were the pioneers in developing dynamic channel-wise topologies. We adopted their basic idea and implemented a graph auto-encoder network and proposed a GA-GCN model.

3. Materials and Methods

3.1. Datasets

In this study, an existing action recognition dataset named NTU RGB+D 120 [2], which extends NTU RGB+D [1] was used.

NTU RGB+D [1] is a sizable dataset containing 56,880 skeletal action sequences, which can be used for the recognition of human actions. 40 volunteers performed the samples, which were classified into 60 classes. Every sample has a distinct action and is ensured to have a maximum of two participants. The actions of volunteers were simultaneously recorded from different angles using three Microsoft Kinect v2 cameras. The two standards suggested by the authors of this dataset are as follows:

1. Cross-subject (X-sub): 20 individuals provide training data, and the remaining 20 provide testing data.
2. Cross-view (X-view): The testing data are derived from the views of camera 1, while the training data are derived from the views of cameras 2 and 3.

NTU RGB+D 120 [2] expands NTU RGB+Dby adding an extra 57,367 skeletal sequences spanning 60 new action classes to become the largest collection of 3D joint annotations specifically designed for human action recognition. 106 participants performed a total of 113,945 action sequences in 120 classes, which were recorded using three cameras. This dataset has 32 distinct configurations, each of

 doi:10.20944/preprints202401.1998.v1

which corresponds to a certain environment and background. The two standards suggested by the authors of this dataset are as follows:

1. Cross-subject (X-sub): 53 individuals provide training data, and the remaining 53 provide testing data.
2. Cross-setup (X-setup): the 32 setups were divided into odd and even numbers where samples with even-numbered setup IDs provide training data, and the remaining samples with odd-numbered setup IDs provide testing data.

3.2. Preliminaries

In this study, a graph is used to represent the human skeleton data. Joints and bones represent the graph's vertices and edges respectively. An adjacency matrix denoted as $A = (V; E; X)$ is used to represent the graph data, where V_1^N is the set of N vertices and E denotes the set of edges. The adjacency matrix represents the strength of the relationship between v_i and v_j. The feature set of N vertices is denoted as X and represented in a matrix of size R^{NxC} and v_i's feature is denoted as $x_i \in R^C$. The following formula is used to obtain the graph convolution:

$$X^{out} = \sum_{v=0}^{N} \mathbf{W} X_j a_{ij} \tag{1}$$

Equation (1) defines the output of the relevant features X^{out} based on the weight W and adjacency matrix A.

The graph autoencoder network is composed of two parts, i.e. the encoder and the decoder. Equation (2) defines the input of the layers X and pooling function $pool()$ of the encoder.

$$X_i = pool(X_{i-1}) \tag{2}$$

Equation (3) defines the input of the layers X, Unpooling function $Unpool()$, and the decode function $decode()$ of the decoder.

$$X_i = decode(X_{N-i}, Unpool(X_{i-1})) \tag{3}$$

Data related to skeletons can be gathered using devices to capture the motion or pose estimation techniques from recorded videos. Video data are often presented as a series of frames, with every frame containing the coordinates of a set of joints. A spatiotemporal graph was constructed by representing the joints as graph vertices and utilizing the inherent connections in the human body structure and time as graph edges using 2D or 3D coordinate sequences for body joints representation. The inputs to GA-GCN are the coordinate vectors of joints located at the graph nodes. This can be compared to image-based CNNs, in which the input is composed of pixel intensity vectors that are located on a 2D image matrix. The input data were subjected to several spatiotemporal graph convolution layers, resulting in the generation of more advanced feature mappings on the graph. The basic SoftMax classifier was subsequently allocated to the matching action class. The entire model was trained using an end-to-end method via back propagation. The proposed pipeline is shown in Figure 2.

3.3. Spatiotemporal Graph Autoencoder Network for Skeleton-based Human Action Recognition Algorithm

In this study, a potent spatiotemporal graph autoencoder network named GA-GCN was built for human action recognition based on skeleton data. We chose to use the graph based on the whole human skeleton as the nature of each joint because previous research has shown it to be more efficient for this task [21,30]. The autoencoder network is composed of 10 fundamental blocks divided into decoder and encoder parts. Subsequently, a global average pooling layer and a softmax classifier are used in order to predict action labels. A pooling layer was added after each encoder block to reduce the total number of joints in half. Each block of the decoder is preceded by an unpooling

 doi:10.20944/preprints202401.1998.v1

layer to double the joints. The number of input channels and joints in the autoencoder blocks are (64,25)-(64,13)-(64,7)-(64,4)-(128,2)-(64,2)-(64,4)-(256,7)-(192,13)-(256,25). Strided temporal convolution reduced the temporal dimensions by half in the fifth and eighth blocks. The network pipeline of the proposed GA-GCN model is illustrated in Figure 2.

3.4. Spatiotemporal Input Representations

Spatiotemporal representations typically refer to data or information that captures the spatial as well as temporal dimensions.

The input of spatiotemporal representations consists of video data represented by skeletal sequences. Resizing process was applied to each sample, resulting in a total of 64 frames.

3.5. Modalities of GA-GCN

The data from eight different modalities: joint, joint motion, bone, bone motion, joint fast motion, joint motion fast motion, bone fast motion, and bone motion fast motion were combined. Table 1 lists the configuration used for each modality. Basically, we change the values of the three variables to obtain eight different modalities in the following order: bone, vel, and fast-motion.

Firstly, the data are the same as the values of joints in a frame for all the frames from the dataset; then, if the bone flag is true then the values of data for joints are updated to the difference between the values of bone pairs in each frame. Then, if the vel flag is true the values of data for joints are updated in the current frame to the difference between the same values of joints in the next frame and current frame. Finally, if the fast-motion flag is true the values of data for joints are updated to the average of the values from the previous, current and next frame.

Table 1. Different modalities configuration flags used in the training process

Modality	bone	vel	fast-motion
joint	FALSE	FALSE	FALSE
joint motion	FALSE	TRUE	FALSE
bone	TRUE	FALSE	FALSE
bone motion	TRUE	TRUE	FALSE
joint fast motion	FALSE	FALSE	TRUE
joint motion fast motion	FALSE	TRUE	TRUE
bone fast motion	TRUE	FALSE	TRUE
bone motion fast motion	TRUE	TRUE	TRUE

4. Results

In this section, the experimental findings are provided.

4.1. Implementation Details

All experiments were carried out using the PyTorch framework for deep learning with a single NVIDIA A100 Tensor Core GPU. Our models were trained using Stochastic Gradient Descent (SGD) which has a momentum value of 0.9 and value of 0.0004 for weight decay. To improve the the training process's stability, a warming strategy was implemented during the initial five epochs as outlined in the study carried out by He et al. [31]. Additionally, the training epoch was set to 65. The learning rate is fixed at 0.1 and decreases by 0.1 at epochs 35 and 55. The resizing of each sample to 64 frames was conducted for both NTU RGB+D and NTU RGB+D 120 datasets. Additionally, the data pre-processing method described by Zhang et al. [28] was employed.

 doi:10.20944/preprints202401.1998.v1

4.2. Experimental results

The summary of the experimental findings of our GA-GCN model on the NTU RGB+D dataset with cross-subject and cross-view is shown in Table 3. The summary of the experimental results of our GA-GCN model on the NTU RGB+D 120 dataset with cross-subject and cross-set is shown in Table 4. The improvement after adding four more modalities and ensemble the results when conducting the experiment on NTU RGB+D with cross-view is shown in Table 2.

Table 2. Comparing of accuracies when ensemble the modalities and add four more modalities to GA-GCN for cross-view on NTU RGB+D experiment

Methods	Accuracy (%)
GA-GCN joint modality	95.14
GA-GCN joint motion modality	93.05
GA-GCN bone modality	94.77
GA-GCN bone motion modality	91.99
GA-GCN after ensemble joint, joint motion, bone and bone motion modalities in our machine	96.51
GA-GCN joint fast motion modality	94.63
GA-GCN joint motion fast motion modality	92.61
GA-GCN bone fast motion modality	94.41
GA-GCN bone motion fast motion modality	91.54
GA-GCN after ensemble joint fast motion, joint motion fast motion, bone fast motion and bone motion fast motion modalities in our machine	96.36
GA-GCN with 8 modalities	
joint, joint motion, bone, bone motion, joint fast motion, joint motion fast motion, bone fast motion and bone motion fast motion	96.7

Table 3. Comparative analysis of classification accuracy with state-of-the-art methods on the NTU RGB+D dataset

Methods	NTU-RGB+D	
	X-Sub (%)	X-View (%)
Ind-RNN [32]	81.8	88.0
HCN [33]	86.5	91.1
ST-GCN [24]	81.5	88.3
2s-AGCN [21]	88.5	95.1
SGN [28]	89.0	94.5
AGC-LSTM [34]	89.2	95.0
DGNN [35]	89.9	96.1
Shift-GCN [36]	90.7	96.5
DC-GCN+ADG [29]	90.8	96.6
PA-ResGCN-B19 [37]	90.9	96.0
DDGCN [38]	91.1	**97.1**
Dynamic GCN [25]	91.5	96.0
MS-G3D [20]	91.5	96.2
CTR-GCN [11]	92.4	96.8
DSTA-Net [39]	91.5	96.4
ST-TR [40]	89.9	96.1
4s-MST-GCN [41]	91.5	96.6
PSUMNet [42]	**92.9**	96.7
GA-GCN	92.3	96.7

 doi:10.20944/preprints202401.1998.v1

Table 4. Comparative analysis of classification accuracy with state-of-the-art methods on the NTU RGB+D 120 dataset

Methods	NTU-RGB+D 120	
	X-Sub (%)	X-Set (%)
ST-LSTM [43]	55.7	57.9
GCA-LSTM [8]	61.2	63.3
RotClips+MTCNN [44]	62.2	61.8
ST-GCN [24]	70.7	73.2
SGN [28]	79.2	81.5
2s-AGCN [21]	82.9	84.9
Shift-GCN [36]	85.9	87.6
DC-GCN+ADG [29]	86.5	88.1
MS-G3D [20]	86.9	88.4
PA-ResGCN-B19 [37]	87.3	88.3
Dynamic GCN [25]	87.3	88.6
CTR-GCN [11]	88.9	90.6
DSTA-Net [39]	86.6	89.0
ST-TR [40]	82.7	84.7
4s-MST-GCN [41]	87.5	88.8
PSUMNet [42]	**89.4**	**90.6**
GA-GCN	88.8	90.4

5. Discussion

This section presents the details of the carried out ablation studies to demonstrate the effectiveness of the proposed spatiotemporal graph autoencoder convolutional network GA-GCN are described. Then the GA-GCN proposed in this study is compared with other cutting-edge methods based on evaluation using two datasets.

The efficacy of the GA-GCN was assessed using ST-GCN [24] as the baseline method that falls under static topology shared graph convolution. To ensure a fair comparison, residual connections were incorporated into ST-GCN as the fundamental building units and substituting its temporal convolution with the temporal modeling module as outlined in Section 3.

5.1. Comparison of GA-GCN Modalities

Table 2 shows how the accuracy increased when the four modalities of joint, joint motion, bone, and bone motion were ensembled when compared to the accuracy of a single modality. Subsequently, four more modalities were added and it was noted that the accuracy further increased compared to the accuracy when just four modalities are used. The eight modalities are as follows: joint, joint motion, bone, bone motion, joint fast motion, joint motion fast motion, bone fast motion, and bone motion fast motion. The experiments' findings indicate that the accuracy of our model increased by 1.0% when the results of four more modalities involving fast motion were ensembled compared with the accuracy obtained with the original four modalities.

5.2. Comparison with the State-of-the-Art

Multi-stream fusion frameworks have been used by many state-of-the-art techniques. To allow for a fair comparison, the same framework as in [25,36] were used for comparison. In particular, data from eight different modalities: joint, joint motion, bone, bone motion, joint fast motion, joint motion fast motion, bone fast motion, and bone motion fast motion were combined for comparison purposes. In Tables 3 and 4, the model developed in this study is compared to state-of-the-art methods evaluated

 doi:10.20944/preprints202401.1998.v1

based on the NTU RGB+D and NTU RGB+D 120 datasets, respectively. Our method outperforms most of the existing state-of-the-art methods when evaluated based on two common datasets.

6. Conclusions and Future Work

In this study, we proposed a novel skeleton-based human action recognition algorithm, named GA-GCN. Our algorithm utilizes the power of the spatiotemporal graph autoencoder network to achieve high accuracy. In comparison to other graph convolutions, GA-GCN has a greater representation capability. Moreover, we added four input modalities to further improve the performance; see Section 4. The GA-GCN was evaluated on two common datasets NTU RGB+D and NTU RGB+D 120 and outperformed most of the existing state-of-the-art methods. Additional experiments on more datasets can be considered as a future work. Furthermore, extra graph edges can be added between significant nodes for specific actions to improve human action recognition performance.

Author Contributions: Conceptualization, H.A. and A.E.; methodology, H.A., A.E., A.M. and F.A.; software, H.A.; validation, H.A.; formal analysis, H.A., A.E., A.M. and F.A.; investigation, H.A., A.E., A.M. and F.A.; resources, A.E, A.M. and F.A.; data curation, H.A.; writing—original draft preparation, H.A.; writing—review and editing, H.A., A.E., A.M. and F.A.; supervision, A.E., A.M. and F.A.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset was downloaded from: https://rose1.ntu.edu.sg/dataset/actionRecognition/.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. IEEE Conference on Computer Vision and Pattern Recognition, 2016.
2. 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 Transactions on Pattern Analysis and Machine Intelligence* **2019**. doi:10.1109/TPAMI.2019.2916873.
3. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; others. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* **2017**.
4. Liu, J.; Shahroudy, A.; Wang, G.; Duan, L.Y.; Kot, A.C. Skeleton-based online action prediction using scale selection network. *IEEE transactions on pattern analysis and machine intelligence* **2019**, *42*, 1453–1467.
5. Johansson, G. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics* **1973**, *14*, 201–211.
6. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595–3603.
7. Liu, J.; Shahroudy, A.; Xu, D.; Kot, A.C.; Wang, G. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE transactions on pattern analysis and machine intelligence* **2017**, *40*, 3007–3021.
8. Liu, J.; Wang, G.; Duan, L.Y.; Abdiyeva, K.; Kot, A.C. Skeleton-based human action recognition with global context-aware attention LSTM networks. *IEEE Transactions on Image Processing* **2017**, *27*, 1586–1599.
9. Hamilton, W.; Ying, Z.; Leskovec, J. Inductive representation learning on large graphs. *Advances in neural information processing systems* **2017**, *30*.
10. Kipf, T.N.; Welling, M. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* **2016**.

11. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13359–13368.
12. Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.J.; Yuan, J.; Thalmann, N.M. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2272–2281.
13. Malik, J.; Elhayek, A.; Guha, S.; Ahmed, S.; Gillani, A.; Stricker, D. DeepAirSig: End-to-End Deep Learning Based in-Air Signature Verification. *IEEE Access* **2020**, *8*, 195832–195843.
14. Bruna, J.; Zaremba, W.; Szlam, A.; Lecun, Y. Spectral Networks and Locally Connected Networks on Graphs. International Conference on Learning Representations (ICLR); Apr 14–16; Banff, AB, Canada, 2014.
15. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems* **2016**, *29*.
16. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR), 2017.
17. Duvenaud, D.K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R.P. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems* **2015**, *28*.
18. Niepert, M.; Ahmed, M.; Kutzkov, K. Learning convolutional neural networks for graphs. International conference on machine learning. PMLR, 2016, pp. 2014–2023.
19. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. International Conference on Learning Representations (ICLR), 2018.
20. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 143–152.
21. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12026–12035.
22. Tang, Y.; Tian, Y.; Lu, J.; Li, P.; Zhou, J. Deep progressive reinforcement learning for skeleton-based action recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5323–5332.
23. Veeriah, V.; Zhuang, N.; Qi, G.J. Differential recurrent neural networks for action recognition. Proceedings of the IEEE international conference on computer vision, 2015, pp. 4041–4049.
24. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. Thirty-second AAAI conference on artificial intelligence, 2018.
25. Ye, F.; Pu, S.; Zhong, Q.; Li, C.; Xie, D.; Tang, H. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 55–63.
26. Zhao, R.; Wang, K.; Su, H.; Ji, Q. Bayesian graph convolution lstm for skeleton based action recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6882–6892.
27. Huang, Z.; Shen, X.; Tian, X.; Li, H.; Huang, J.; Hua, X.S. Spatio-temporal inception graph convolutional networks for skeleton-based action recognition. Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2122–2130.
28. Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; Zheng, N. Semantics-guided neural networks for efficient skeleton-based human action recognition. proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1112–1121.
29. Cheng, K.; Zhang, Y.; Cao, C.; Shi, L.; Cheng, J.; Lu, H. Decoupling gcn with dropgraph module for skeleton-based action recognition. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. Springer, 2020, pp. 536–553.
30. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic convolution: Attention over convolution kernels. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11030–11039.

31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
32. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5457–5466.
33. Li, C.; Zhong, Q.; Xie, D.; Pu, S. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055* **2018**.
34. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1227–1236.
35. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with directed graph neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 7912–7921.
36. Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; Lu, H. Skeleton-based action recognition with shift graph convolutional network. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 183–192.
37. Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1625–1633.
38. Korban, M.; Li, X. Ddgcn: A dynamic directed graph convolutional network for action recognition. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. Springer, 2020, pp. 761–776.
39. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. Proceedings of the Asian Conference on Computer Vision, 2020.
40. Plizzari, C.; Cannici, M.; Matteucci, M. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding* **2021**, *208*, 103219.
41. Chen, Z.; Li, S.; Yang, B.; Li, Q.; Liu, H. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. Proceedings of the AAAI conference on artificial intelligence, 2021, pp. 1113–1122.
42. Trivedi, N.; Sarvadevabhatla, R.K. PSUMNet: Unified Modality Part Streams are All You Need for Efficient Pose-based Action Recognition. Computer Vision–ECCV 2022 Workshops: Proceedings, Part V. Springer, 2023, pp. 211–227.
43. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer, 2016, pp. 816–833.
44. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. Learning clip representations for skeleton-based 3d action recognition. *IEEE Transactions on Image Processing* **2018**, *27*, 2842–2855.