Spatiotemporal Graph Autoencoder Network for Skeleton-Based Human Action Recognition

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

Posted Date: 26 July 2024

doi: 10.20944/preprints202401.1998.v3

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

Abstract: The task of human action recognition (HAR) based on skeleton data is a challenging yet crucial technique owing to its wide-ranging applications in numerous domains, including patient monitoring, security surveillance, and observation of human-machine interactions. While numerous algorithms have been proposed in an attempt to distinguish between a myriad of activities, most practical applications necessitate highly accurate detection of specific activity types. This study proposes a novel and highly accurate spatiotemporal graph autoencoder network for HAR based on skeleton data. Furthermore, an extensive investigation was conducted employing diverse modalities. To this end, a spatiotemporal graph autoencoder was constructed to automatically learn both spatial and temporal patterns from human skeleton datasets. The powerful graph convolutional network, designated as GA-GCN, developed in this study, notably outperforms the majority of existing state-of-the-art methods when evaluated on two common datasets, namely NTU RGB+D and NTU RGB+D 120. On the first dataset, the proposed approach achieved accuracies of 92.3% and 96.8% for the cross-subject and cross-view evaluations, respectively. On the more challenging NTU RGB+D 120 dataset, GA-GCN attained accuracies of 88.8% and 90.4% for the cross-subject and cross-set evaluations, respectively.

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

1. Introduction

The recognition and analysis of human actions is a critical subfield within the domains of computer vision and deep learning, with the primary objective of automatically detecting and classifying human actions or gestures from video data [1–3]. Sophisticated algorithms and models capable of understanding and interpreting the dynamics of human movements are essential for this purpose [4,5]. The recognition and interpretation of human actions play a crucial role in various practical applications, such as video surveillance, healthcare systems, robotics, human-computer interaction, and beyond [6–8]. The ability to extract meaningful information from video sequences enables machines to comprehend and respond to human actions, thereby enhancing the efficiency and safety of many domains [9,10].

This emerging field of research leverages the capabilities of deep learning techniques to capture the temporal and spatial features inherent in video data [11,12]. Recent advancements in the use of 3D graph convolutional neural networks (3D GCNs) have further improved the accuracy of action recognition models [13,14]. Notable datasets, including NTU RGB+D [1], NTU RGB+D 120 [2], NW-UCLA, and Kinetics [3], have become established benchmarks for evaluating and analyzing the performance of these techniques, driving further research and innovation in the field.

1.1. Human Action Recognition Based on Skeleton Data

The primary objective of employing action prediction algorithms is to anticipate the classification label of a continuous action based on a partial observation along the temporal dimension [4]. This task of predicting human activities prior to their full execution is regarded as a subfield within the broader scientific area of human activity analysis. This field has garnered considerable scholarly

 doi:10.20944/preprints202401.1998.v3

interest due to its extensive array of practical applications across domains such as security surveillance, the observation of human-machine interactions, and medical monitoring [4].

The ability to accurately forecast the unfolding of human actions based on incomplete sensory data holds immense potential for enhancing the capabilities of intelligent systems. By proactively recognizing ongoing activities, predictive HAR models can enable preemptive responses, improved decision-making, and more seamless human-machine coordination. This predictive capability is particularly valuable in time-critical applications, where the early identification of actions can contribute to enhanced safety, efficiency, and situational awareness [4].

Extant biological research has demonstrated that human skeleton data, as depicted in Figure 1, are sufficiently informative to represent human behavior, despite the absence of appearance-related information [5]. This finding is grounded in the inherently three-dimensional spatial context within which human activities are conducted, rendering three-dimensional skeletal data an appropriate means of capturing human activity dynamics [5]. The efficient and convenient real-time acquisition of 3D skeletal information can be achieved through the utilization of affordable depth sensors, such as Microsoft Kinect and Asus Xtion. Consequently, the analysis of 3D skeleton data has become a prominent area of scholarly study within the field of human activity recognition [6–8].

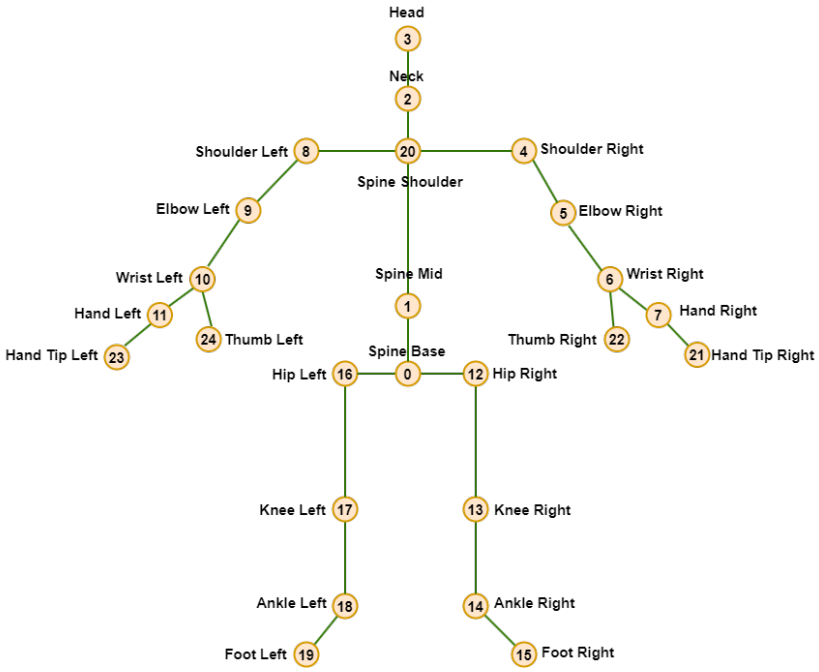

Figure 1. Human skeleton with circles representing joints and lines representing bones.

The utilization of 3D skeleton data for activity analysis offers several key advantages, including its conciseness, sophisticated representational capacity, and resilience to variations in viewpoint, illumination, and surrounding visual distractions [5–8]. These properties have contributed to the growing prominence of skeleton-based approaches within the broader landscape of human action recognition research. Continued advancements in 3D skeletal data acquisition and modeling techniques hold significant promise for further enhancing the accuracy and robustness of activity recognition systems across a wide range of applications.

1.2. 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

 doi:10.20944/preprints202401.1998.v3

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 continue to evolve, offering promising avenues for further research and development.

1.3. Proposed Spatiotemporal Model for Human Action Recognition

The primary objective of this study was to develop a novel and highly accurate algorithm for human motion detection and action recognition. To this end, we focused on leveraging special practical scenarios and leveraging the latest deep learning (DL) technologies, such as graph convolutional networks (GCNs), autoencoders, and one-class classifiers, to construct a highly accurate human action recognition (HAR) framework.

The study was conducted using the well-known NTU RGB+D 120 dataset [2], which is an extension of the NTU RGB+D dataset [1] and provides extracted skeletal motion data for 120 distinct motion classes. This dataset was selected due to its comprehensive coverage of human activities and its widespread adoption within the HAR research community.

The core of the proposed approach is a spatiotemporal graph autoencoder (GA-GCN) network architecture, as illustrated in Figure 2. The input to the network is a spatiotemporal graph constructed from the skeletal sequence data, as described in [12]. This graph-structured representation is then fed into the autoencoder for unsupervised learning of the spatial and temporal patterns inherent in the data.

The GA-GCN architecture is based on the channel-wise topology refinement GCN (CTR-GCN) model proposed by Chen et al. [11], which has demonstrated strong performance on various graph-based tasks. To further enhance the learning process, we incorporated additional skip connections to enable information flow from the decoder layers to the encoder layers, thereby facilitating more effective spatiotemporal feature extraction and reconstruction.

The verification process involves the application of thresholding to the reconstruction loss, as described in [13], to identify abnormal or anomalous human activities within the input skeletal data. This approach leverages the autoencoder's ability to learn the normative patterns of human motion, allowing for the detection of deviations from these learned representations. The key contributions of this work are threefold:

1. The development of a novel spatiotemporal graph-autoencoder network for skeleton-based HAR, which effectively captures the complex spatial and temporal dynamics of human movements, offering a significant advancement in feature extraction and representation.
2. Outperforming most existing methods on two widely used skeleton-based HAR datasets.
3. Achieving notable performance improvements by incorporating additional modalities, as demonstrated in the experimental evaluation presented in Section 4.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2024 doi:10.20944/preprints202401.1998.v3

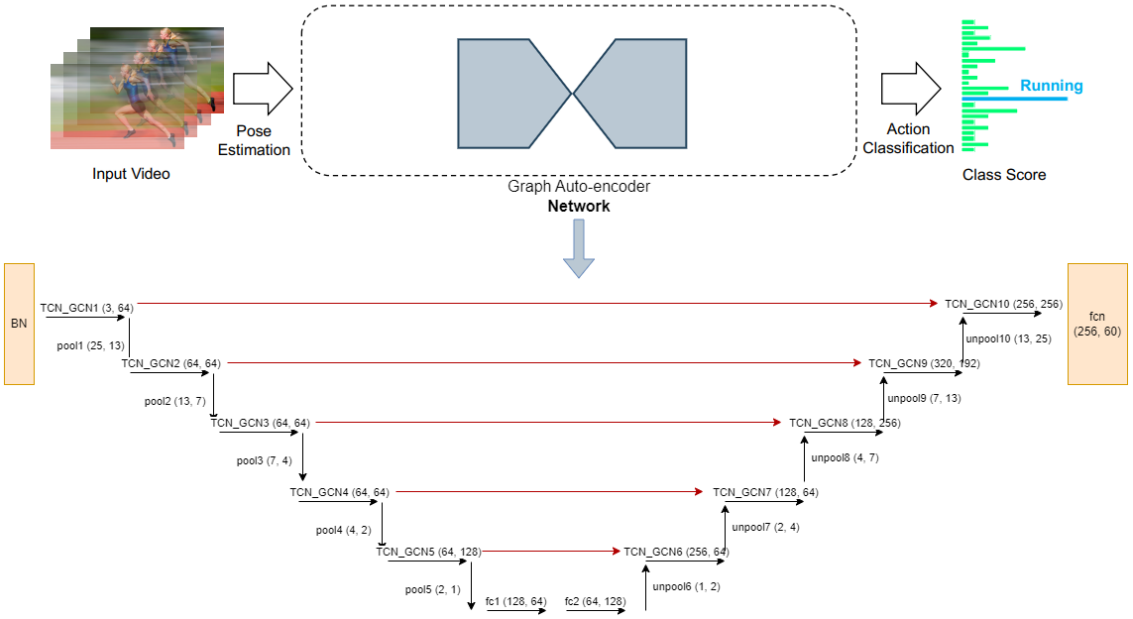

Figure 2. Network architecture schematic overview of the proposed spatiotemporal graph autoencoder network for HAR based on skeleton data. The top figure is an overview of the entire pipeline, and the lower figure shows the input channels and joints of the graph autoencoder parts and provides details 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 (GCNs)

CNN-based methods have achieved great results in RGB-based processing compared to skeleton-based representations. However, GCN-based methods overcome CNN weaknesses for processing skeleton data. 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] has been successfully adapted for skeleton-based action recognition [20–26]. GCN-based techniques emphasize skeleton graph modeling due to the significance of graphs in GCNs. Based on the topological variations, GCN-based techniques can be categorized into the following:

 doi:10.20944/preprints202401.1998.v3

1. Static and dynamic techniques: In static techniques, the topologies of GCNs remain constant throughout the inference process, whereas they are dynamically inferred during 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 different channels or channel groups in topology non-shared techniques.

In the context of static approaches, Yan et al. [24] proposed a Spatial-Temporal GCN (ST-GCN) that predefined a fixed topology based on the human body structure. Incorporating multi-scale graph topologies into GCNs was proposed by Liu et al. [20] and Huang et al. [27] to facilitate the modeling of joint relationships across different ranges.

Regarding dynamic approaches, Li et al. [6] suggested an A-links inference component that records action-related correlations. Self-attention methods enhanced topology learning by modeling the correlation between each pair of joints [21,28]. These techniques use regional information to infer the relationships between two joints. Ye et al. [25] proposed a dynamic GCN method that learns correlations between joint pairs by incorporating contextual data from joints. Dynamic methods offer greater generalization capabilities than static methods due to their dynamic topologies.

In terms of topology sharing, 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 with identical topology. Most GCN-based techniques, including the aforementioned static [20,24,27] and dynamic [6,21,25,28] methods, operate in a topology shared manner.

Topology non-shared techniques naturally overcome the drawbacks of topology shared techniques by employing various topologies in different channels or channel groups. Decoupling Graph Convolutional Network (DC-GCN) used various graph representations for different channels [29], but it encountered optimization challenges when constructing channel-wise topologies due to the large number of parameters. In skeleton-based Human Action Recognition (HAR), topology non-shared graph convolutions have rarely been investigated. Chen et al. [11] were the pioneers in developing dynamic channel-wise topologies. This study adopted their basic idea and implemented a graph auto-encoder network, proposing a Graph Auto-Encoder GCN (GA-GCN) model.

3. Materials and Methods

3.1. Datasets

This study utilized an existing action recognition dataset known as NTU RGB+D 120 [2], which extends the NTU RGB+D dataset [1]. NTU RGB+D [1] is a sizable dataset containing 56,880 skeletal action sequences, designed for the recognition of human actions. The dataset was collected from 40 volunteers performing samples across 60 action categories. Each video sample contains a single action and involves a maximum of two participants. The volunteer actions were simultaneously recorded using three Microsoft Kinect v2 cameras from various viewpoints. The dataset authors suggested two standard evaluation protocols:

1. The subject samples are divided into two halves: 20 individuals provide training samples, while the remaining 20 provide testing samples. This standard is named cross-subject (x-sub).
2. The testing samples are derived from the views of camera 1, while the training samples are derived from the views of cameras 2 and 3. This standard is named cross-view (x-view).

The NTU RGB+D 120 dataset [1] expands upon NTU RGB+D by adding 57,367 skeletal sequences spanning 60 new action classes, making it the largest collection of 3D joint annotations designed for human action recognition (HAR). The dataset was collected from 106 participants performing a total of 113,945 action sequences in 120 classes, recorded using three cameras. The dataset has 32 distinct configurations, each corresponding to a specific environment and background. The dataset authors suggested two standard evaluation protocols:

 doi:10.20944/preprints202401.1998.v3

1. The subject samples are divided into two halves: 53 individuals provide training data, and the remaining 53 provide testing data. This standard is named cross-subject (c-sub).
2. The 32 setups were separated into two halves: sequences with even-numbered setup numbers provide training samples, and the remaining sequences with odd-numbered setup numbers provide testing samples. This standard is named cross-setup (x-setup).

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 (joints) and E denotes the set of edges (edges). The adjacency matrix models the strength of the relationship between v_i and v_j. The input features of N vertices are 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 consists 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, the Unpooling function $Unpool()$, and the decode function $decode()$ of the decoder.

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

Skeleton data can be acquired using motion capture devices or pose estimation techniques from recorded videos. Video data are often presented as a series of frames, with each 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 relationships 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, similar to how image-based CNNs use pixel intensity vectors 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. A basic SoftMax classifier was subsequently used to predict the matching action class. Our proposed model, GA-GCN, was trained using an end-to-end method via back propagation, as shown in Figure 2.

3.3. Spatiotemporal Graph Autoencoder Network for Skeleton-Based HAR Algorithm

This study proposes a potent spatiotemporal graph autoencoder network, named GA-GCN, for skeleton-based HAR. Previous research has shown that using graphs is more efficient for this task [21,30], so we chose to represent the human skeleton as a graph with the nature of each joint. The autoencoder network is composed of 10 fundamental blocks divided into decoder and encoder parts. A global average pooling layer and a softmax classifier are used to predict action labels. A pooling layer was added after each encoder block to minimize the overall number of joints by half, and each block of the decoder is preceded by an unpooling 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. Figure 2 demonstrates the proposed network pipeline, GA-GCN.

 doi:10.20944/preprints202401.1998.v3

3.4. Spatiotemporal Input Representations

Spatiotemporal representations refer to data or information that reflects the spatial and temporal dimensions. The input of spatiotemporal representations consists of video data represented by skeletal sequences. A 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, the values of the three variables (bone, vel, and fast-motion) were changed to obtain eight different modalities.

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

This section details the implementation specifics and the empirical findings of the study.

4.1. Implementation Details

The experiments were conducted using the PyTorch framework and executed on a single NVIDIA A100 Tensor Core GPU. The models were trained using Stochastic Gradient Descent (SGD) with a momentum value of 0.9 and a weight decay value of 0.0004. To enhance the stability of the training process, a warming strategy was implemented during the initial five epochs, as outlined in the study by He et al. [31]. Furthermore, the model was trained for 65 epochs, with the learning rate decreased by 0.1 at epochs 35 and 55. For both the NTU RGB+D and NTU RGB+D 120 datasets, each sample was resized to 64 frames. Additionally, the data pre-processing method described by Zhang et al. [28] was employed.

4.2. Experimental Results

The performance of the GA-GCN model on the NTU RGB+D 60 dataset is illustrated through the confusion matrix, precision, recall, F1-score, and specificity metrics. Figure 3 shows the confusion matrix for the cross-view evaluation, detailing the model's ability to classify actions correctly across different camera views. Figure 4 illustrates these metrics for the cross-view evaluation, each bar represents a distinct action class, illustrating the model's performance in terms of precision (positive predictive value), recall (sensitivity), the F1-score (harmonic mean of precision and recall), and specificity (negative predictive value). Figure 5 presents the confusion matrix for the cross-subject

 doi:10.20944/preprints202401.1998.v3

evaluation, highlighting the model's performance across different subjects. Figure 6 presents the precision, recall, F1-score and specificity for the cross-subject evaluation, highlighting the model's ability to recognize actions performed by different subjects. These figures reveal areas of high accuracy as well as common misclassifications, providing a comprehensive overview of GA-GCN's effectiveness in human action recognition for all 60 classes.

Our GA-GCN model exhibits notable limitations with specific classes in the NTU RGB+D 60 dataset with cross-view and evaluation, the worst classes in recognition accuracy are Writing, Reading, Type on a Keyboard, Take Off a Shoe, Put On a Shoe and play with phone/tablet. These actions are challenging due to their subtle hand movements and similarities with other activities, leading to lower recognition accuracy. For example, Writing and Reading involve subtle hand gestures that are difficult to distinguish, while Type on a Keyboard shows similar hand movements to other actions. The actions Take Off a Shoe and Put On a Shoe are particularly problematic due to their subtlety and similarity.

Figure 4 displays the precision, recall, and F1-score metrics for the GA-GCN model across all 60 action classes in the NTU RGB+D 60 dataset under the cross-view setting. Each bar represents a distinct action class, illustrating the model's performance in terms of precision (positive predictive value), recall (sensitivity), and the F1-score (harmonic mean of precision and recall).

Table 3 presents the summary of the experimental findings of the proposed GA-GCN model on the NTU RGB+D dataset, evaluated under the cross-subject and cross-view settings. Table 4 summarizes the experimental results of the GA-GCN model on the NTU RGB+D 120 dataset, tested under the cross-subject and cross-set scenarios. The improvement achieved by adding four additional modalities and ensembling the results when conducting the experiment on the NTU RGB+D dataset is shown in Table 2, using the cross-view standard.

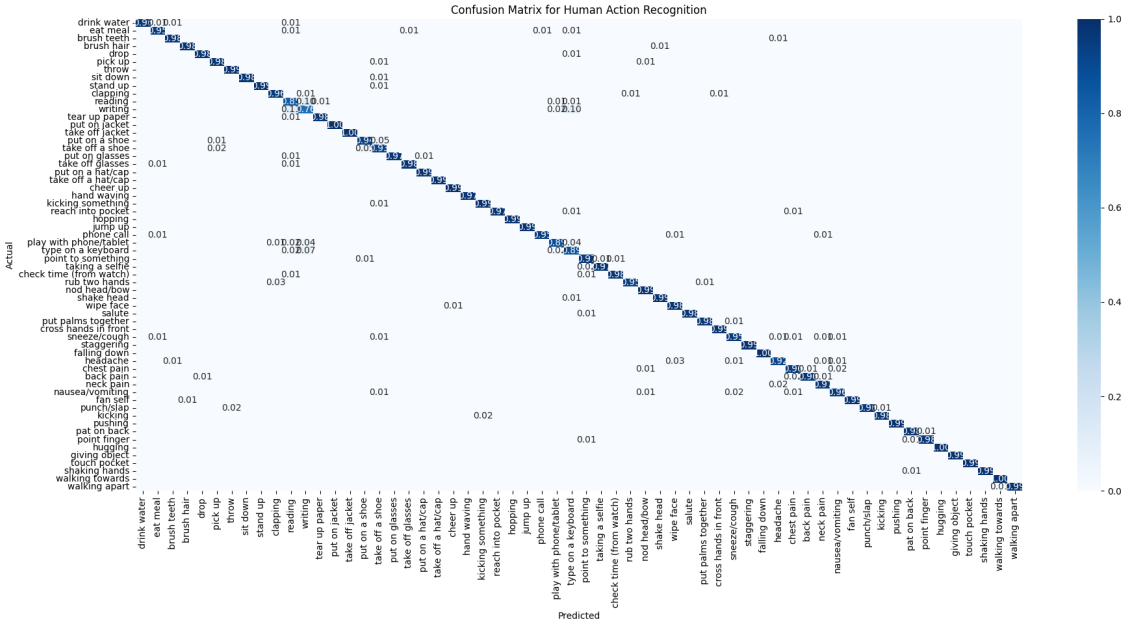

Figure 3. Confusion Matrix of GA-GCN on Cross-View Evaluation of NTU RGB+D 60 Dataset.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2024 doi:10.20944/preprints202401.1998.v3

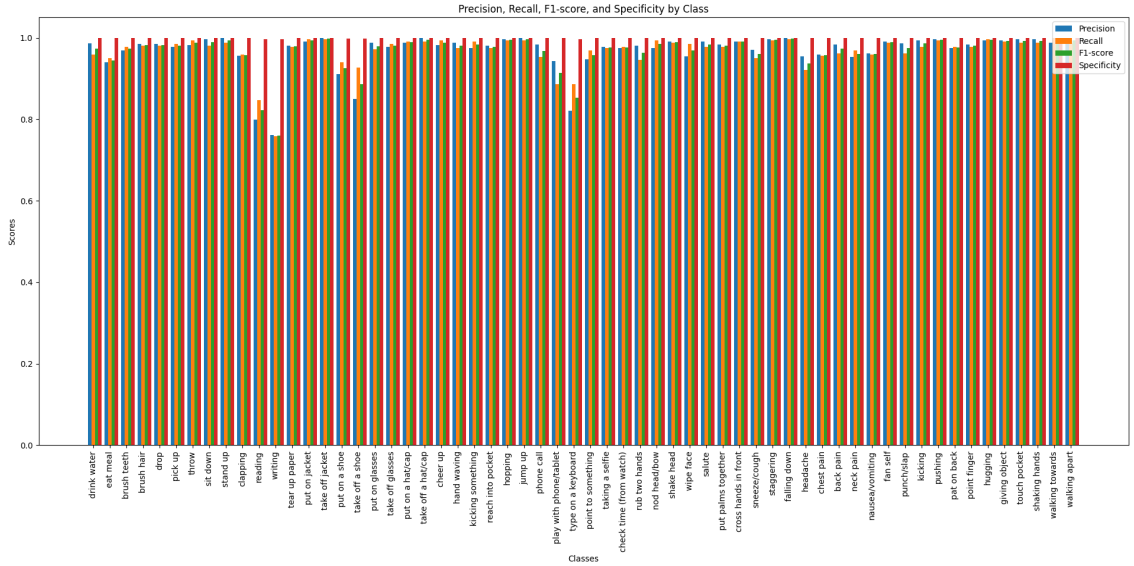

Figure 4. Precision, Recall, F1-Score, and Specificity of GA-GCN on Cross-View Evaluation of NTU RGB+D 60 Dataset.

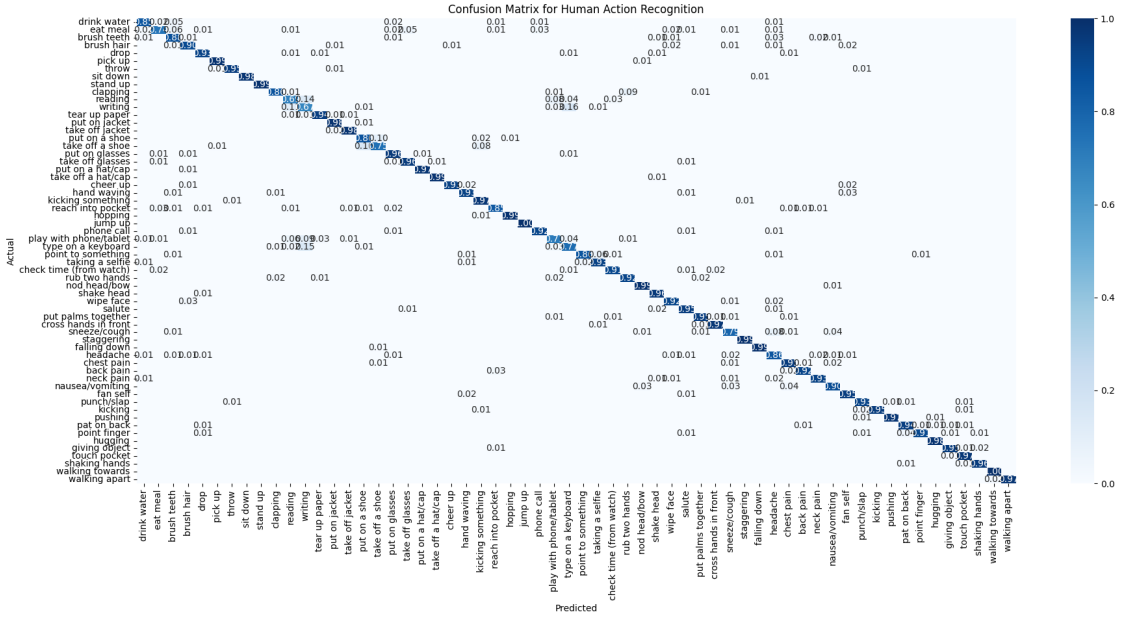

Figure 5. Confusion Matrix of GA-GCN on Cross-Subject Evaluation of NTU RGB+D 60 Dataset.

 doi:10.20944/preprints202401.1998.v3

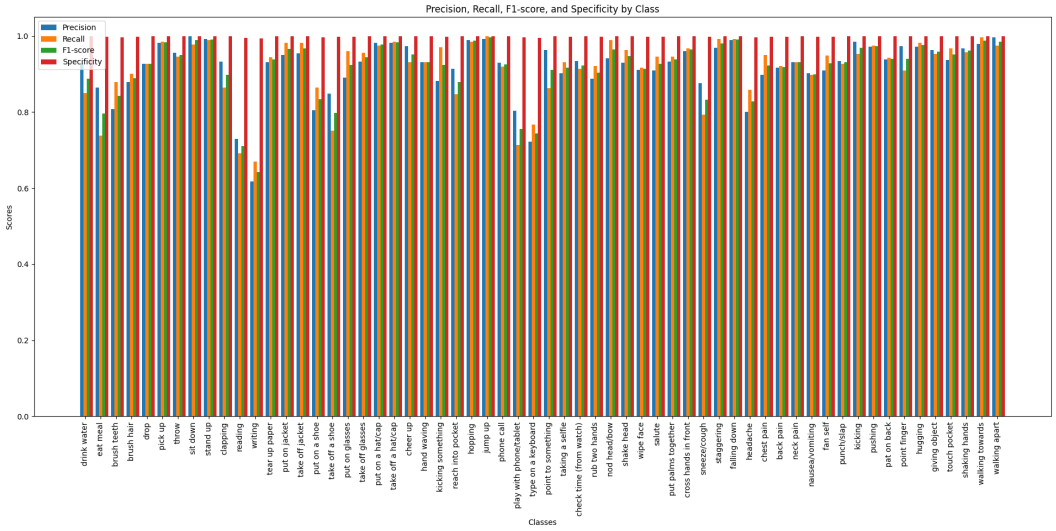

Figure 6. Precision, Recall, F1-Score, and Specificity of GA-GCN on Cross-Subject Evaluation of NTU RGB+D 60 Dataset.

.

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	96.8

Table 3. Comparative analysis of classification accuracy with cutting-edge techniques using NTU RGB+D datase.t

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	96.8

 doi:10.20944/preprints202401.1998.v3

Table 4. Comparative analysis of classification accuracy with cutting-edge techniques using 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	90.5

5. Discussion

This section presents the details of the ablation studies conducted to demonstrate the performance of the proposed spatiotemporal graph autoencoder convolutional network, GA-GCN. Furthermore, the GA-GCN model proposed in this study is compared with other state-of-the-art methods using two benchmark datasets.

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

5.1. Comparison of GA-GCN Modalities

Table 2 shows how the accuracy increased when the results of the four modalities (joint, joint motion, bone, and bone motion) were ensembled, compared to the accuracy of a single modality. Subsequently, four more modalities were added, and it was observed that the accuracy further increased compared to the accuracy when just the original four modalities were 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 experimental findings indicate that the accuracy of the model increased by 1.0% when the results of the four additional modalities involving fast motion were ensembled, compared to the accuracy obtained with the original four modalities.

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

Many state-of-the-art techniques have employed multimodality fusion frameworks. To ensure a fair comparison, the same structure as in [25,36] was used. In particular, data from the eight different modalities mentioned above were combined for comparison purposes. In Tables 3 and 4, the model developed in this study is compared to cutting-edge techniques evaluated using the NTU RGB+D and NTU RGB+D 120 datasets, respectively. The proposed technique outperforms most of the existing state-of-the-art methods when evaluated based on these two common datasets.

6. Conclusions and Future Work

In this work, we proposed a novel skeleton-based human action recognition (HAR) algorithm, named GA-GCN. Our algorithm leverages the power of the spatiotemporal graph autoencoder network to achieve high accuracy. The incorporation of the autoencoder enhances the model's ability to capture complex spatial and temporal dynamics in human movements. Compared to the other graph convolution methods, GA-GCN exhibits a greater representation capability. Furthermore, we added four input modalities to enhance the performance even further, achieving notable performance improvements as demonstrated in the experimental evaluation presented in Section 4. The GA-GCN was

 doi:10.20944/preprints202401.1998.v3

evaluated on two widely used datasets, NTU RGB+D and NTU RGB+D 120, and it outperformed most of the existing state-of-the-art methods. Additional experiments on more datasets can be considered as potential future work. Furthermore, extra graph edges can be added between significant nodes for specific actions to improve HAR 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*.

 doi:10.20944/preprints202401.1998.v3

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.

 doi:10.20944/preprints202401.1998.v3

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.