Proceedings of the 27th International Conference on Automation & Computing, University of the West of England, Bristol, UK, 1-3 September 2022

An Investigation of GCN-based Human Action Recognition Using Skeletal Features

Chuan Dai[1], Yajuan Wei[1,2], Zhijie Xu[1], Minsi Chen[1], Ying Liu[2], Jiulun Fan[2]
[1]School of Computing and Engineering
University of Huddersfield
Huddersfield, UK
[2]School of Computer Science
Xi'an University of Posts and Telecommunications
Xi'an, China
{chuan.dai, yajuan.wei, z.xu, m.chen}@hud.ac.uk, ly_yolanda@sina.com, jiulunf@xupt.edu.cn

2022 27th International Conference on Automation and Computing (ICAC) | 978-1-6654-9807-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAC55051.2022.9911078

Abstract— **Human action recognition is one of the most challenging and attractive areas in the field of computer vision. Conventional research on human action recognition has mainly focused on data modality of video or optical flow. However, the human skeletal feature has much stronger expressive power of motion dynamics, which is not sensitive to illumination and scene variation. Owing to the advantages of deep learning approaches on skeleton data in recent years, many pilot approaches have been proposed, which are merited by their significant performance enhancements on both baseline and large-scale datasets. This research investigates these models and their breakthroughs, especially focusing on the graph convolution network (GCN) and skeleton-based data techniques. The report work mainly covers the following aspects: comparing RNN, CNN and GCN-based approaches from the perspective of their operational logics; a detailed review of the best referred models in recent years; a development framework of skeletal feature-based human action recognition framework is proposed with preliminary assessments using benchmarking datasets; and finally, the envisaged future directions for skeletal feature-based human action recognition study are discussed.**

Keywords: Human Action Recognition; Skeleton Data; GCN

I. Introduction

Human action recognition refers to the process of identifying and understanding human actions and is critical for series of real-world applications, for example, it can be used in autonomous navigation system [1] to ensure the safety of road traffic, and for video surveillance [2], where dangerous human behaviors can be detected. Other applications have also included human-computer interaction [3], entertainment [4], and video retrieval [5].

In general, raw input data for human action recognition includes several modalities and can be divided into visual modalities and non-visual modalities [6]. This study focuses on skeleton-based data, which falls into the category of visual modalities. Skeleton data is encoded from the trajectories of human body joints for both intra-frame and inter-frame joints. Other objects and background information in frames are ignored. Therefore, such data can effectively represent the characteristics and various changing trends of human behaviors. Skeleton data can be mainly acquired by applying video-based pose estimation algorithms [7] or depth sensors [8].

Skeleton data is becoming more important than other data modalities because of the following aspects: 1) For spatial information, the position relationship of intra-frame joints plays an important role in modeling the intrinsic relation of structural information [9][10]. 2) For temporal information, different positions of the same joint show strong temporal relationship with an inter-frame manner [9][10]. 3) For the co-occurrence relationship, once spatial and temporal information is applied for learning and modeling simultaneously, the co-occurrence relationship is reflected [9]. 4) Compared to the most used RGB and optical flow, skeleton data has a small total amount of data, so the computational burden is correspondingly small [11]. 5) Skeleton data is not sensitive to background noise and illumination changes [11].

The main contributions of this study can be summarized as follows: 1) From the point of skeleton-based human action recognition, this study compares RNN, CNN and GCN approaches from the perspective of development logic of deep learning principle. 2) The focus of this paper is to give a detailed review of the 11 papers with the highest citation rate in recent years, since a lot of relevant literatures have been reviewed in previous surveys [6][9][10][11][12][13] in recent years. 3) This study, for the first time, presents a development framework of skeleton-based human action recognition approaches, and analyzes the evolution and development logic among each method.

II. Background of GCN

A. Deep Learning-based Action Recognition with Skeleton Data

In RNN-based methods [14][15][16], skeleton sequence was usually constructed as a sequence of coordinate vectors due to the inherent suitability of RNN for handling time series data. However, although the impressive results have been achieved, the biggest drawback is that it ignores the spatial information which always carry strong dependencies in skeleton-based data [11]. Meanwhile, RNN and long-short term memory (LSTM) can usually model the long-term context information in temporal dimension, but it is difficult to complete the modeling of high-level features [17]. Therefore, there are certain limitations in using RNN to deal with skeleton-based data for human action recognition.

To solve the shortcomings of RNN in identifying skeleton-based data, many studies then investigated the possibility of using CNN-based methods [3][18]. The idea of using CNN is based largely on its success in image processing. Firstly, skeleton-based coordinate data is transformed into 2D pseudo images, and then human action recognition is completed by using CNN to identify and predict images. Based on remarkable achievements of CNN to images, the results in skeleton-based data were also sound. However, CNN brings new problems. For example, to obtain 2D pseudo images and encode temporal information, the corresponding computational complexity is significantly increased. In addition, the learning and modeling of long-range relationship is still a problem [10].

Neither RNN nor CNN can completely model the complicated spatial temporal human body skeleton. Since skeleton can be naturally represented as a graph, the joints of human body are represented as vertices, and the bones connecting the joints are illustrated as edges. It's naturally much more expressive than sequence vectors or 2D pseudo images. Therefore, the GCN-based methods have natural advantages in graph structure utilization. It can not only generate skeleton structure of arbitrary form, but also distinguish various information in spatial and temporal domain to the maximum extent. Many studies have consistently proved that GCN is a more suitable method for extracting human body skeleton data than sequence vectors or 2D pseudo images [19][33][34].

B. *Spatial and Spectral Domain GCN*

The principle of building GCN on graph generally follows two flows, spatial-based and spectral-based approaches.

The research of spatial-based approaches started much earlier than spectral-based approaches. It inherits ideas from building RNN on graph to define graph convolutions by information propagation. Spatial-based approaches define convolutions directly on the graph-based topology. Operations such as message-passing and representation methods are used to aggregate information between vertices and make a prediction. The major challenge of spatial-based approaches is defining the convolution operation with differently sized neighborhoods.

Spectral-based approaches have a solid mathematical foundation in graph signal processing. Spectral-based approaches define graph convolutions by introducing filters from the perspective of graph signal processing, where the graph convolutional operation is interpreted as removing noises from graph signals. Operations such as the graph Fourier transform, or its extensions are used to aggregate vertex information and make a prediction. The drawback is that it is computationally expensive since the kernel is defined in Fourier space and the graph Fourier transform is expensive to compute.

III. Skeleton-based Human Action Recognition with GCN

A. *Naive Spatial Temporal GCN*

Wang et al. [20] proposed the idea of representing skeleton data as graphs, which laid a foundation for the subsequent works. Then Yan et al. [19] proposed spatial temporal graph convolution networks (ST-GCN) based on deep learning principle. Before ST-GCN [19] was proposed, the conventional ways for handling skeleton-based data are hand-crafted parts and the mechanism of traversal. The main pain points of those methods are the limitation of expression potency, difficulty in generalization, and low algorithm efficiency. To solve the above problems, ST-GCN [19] first proposes to use spatial temporal method to address human action recognition problems with GCN.

As shown in Fig. 1, each blue vertex represents a joint of human body, and edges are divided into spatial edges and temporal edges. The spatial edges refer to the edges connecting intra-frame vertices, while the temporal edges refer to the connections of the same vertex between consecutive frames. Firstly, the set of V and E are defined, where V is constructed in a single frame, while E is constructed in a single frame and consecutive frames, respectively. Secondly, the network structure and corresponding sampling function and weight function are defined on this basis. In addition, several feasible partitioning strategies are analyzed. Adopting deformable convolutional networks [21] to skeleton data is one of the highlights worth mentioning in this structure, the authors set offset to learn the three types of vertices in partitioning. Finally, the spatial temporal model is proposed.

At the implementation level, the network sets 9 spatial temporal graph convolution layers. To avoid overfitting, data augmentation is applied, which has been verified in related research [22]. In addition, TCN [23] is used as the baseline in this paper, and this method has been applied in a series of subsequent studies [24][25][26][27] as well. The experimental results show that ST-GCN [19] has better performance in both unconstrained and constraint datasets. The two major datasets, Kinetics-Skeleton [28][7] and NTU RGB+D [29], are also utilized as baseline in subsequent studies.

ST-GCN [19] has several outstanding advantages, which play a leading role in subsequent research work.

- It introduces the idea of applying spatial temporal method in GCN to approach the skeleton-based human action recognition.
- This network architecture supports datasets containing different numbers of joint or joint connectivity.
- It is a generic design with automatic learning of spatial and temporal patterns from input data.
- This kind of spatial temporal GCN has excellent expressive power and generalization capability.

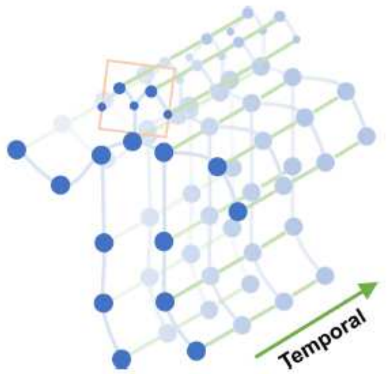

Figure 1. The spatial temporal graph of a skeleton sequence [19]

- Input data is compatible with 2D datasets, such as skeleton data converted by OpenPose [7], and 3D datasets, such as NTU RGB+D [29].
- This spatial temporal GCN can be used as an effective supplementary way to learn human action by RGB or optical flow.

However, as more research being carried out, some drawbacks of ST-GCN [19] have been revealed, namely:

- Drawback 1: It is designed as a fixed graph structure based on the natural physical structure of human skeleton. Therefore, the semantic connections between each joint in each frame are ignored [30][12].
- Drawback 2: The skeleton graph is heuristically predefined, and it represents only the physical structure of the human body. Therefore, it is possibly suboptimal for human action recognition tasks [24][25].
- Drawback 3: The structure of GCN is hierarchical, with its different layers containing multiple levels of semantic information. However, because the topology of the graph is fixed at all layers, which lacks the capability and flexibility to model the multiple layers of semantic information contained in all layers [24][30][31].
- Drawback 4: Using a fixed graph structure may not be the best solution for all samples used to distinguish between different action classes [24].
- Drawback 5: The feature vectors attached to each vertex contain only the 2D or 3D coordinates of the joint that can be considered as first-order information (spatial coordinates of joints and bones) of skeleton data. However, second-order information (bones lengths or directions), representing bone features between a pair of joints, is not explored [24][9].
- Drawback 6: The joint connections in different positions are unbalanced. When torso joints are overly smooth, limb joints may remain in a less smooth state, which is leading to difficulties in sharing features between two limb joints [25].
- Drawback 7: Its convolution kernel can only collect local features, and the skeleton graph of physical connection lacks flexibility, which is not conducive to the distinction between actions [12][27][32][26][33][34].
- Drawback 8: Its design is exceedingly sophisticated, and over parametric, which leads to inefficient model training and inference [12][27].
- Drawback 9: It defines the receptive fields of spatial graph and temporal graph based on intuition. However, the expressive power of this definition is limited [27][35][33][34].

To address the above-mentioned shortcomings, there are 10 key papers reported major improvements on the ST-GCN [19] from different angles, which can be divided into the following categories: devised spatial temporal GCN, two-stream GCN, attention GCN, encoder-decoder GCN, and other Misc. (see Table I).

B. Devised Spatial Temporal GCN

1) NAS-GCN

The core purpose of this model [30] is to automatically design neural network structure for skeleton based GCN. Its main innovation is to combine the skeleton-based GCN problem with neural architecture search (NAS) [36] and adapt it to the skeleton-based problem by adjusting and enhancing the NAS [36] method.

Different from the conventional spatial-based or spectral-based GCN approaches, this work builds a dynamic graph based on various semantic information through the interaction between joints, while other methods compute the importance weight of different representations or frames.

The method proposed in this paper [30] mainly includes the following two points that deserve our attention:

a) For the search space

It utilizes gaussian function used in 2s-AGCN [24] to compute the connection strength between two vertices. In the search space, Chebyshev polynomial functions of different orders are built on various layers, and the network determines the order and polynomial composition of each layer.

Previous NAS [36] reduces computational effort by searching a single block. However, the authors argue that different feature layers contain different levels of semantic content and therefore prefer to use layer-specific mechanisms to build graphs. So, the search is conducted across the entire GCN network instead of each individual block.

b) For the search algorithm

A devised way of combining cross-entropy evolution strategy [37] and importance-mixing (CEIM) is proposed. It applies gaussian distribution to model the architecture and updates the search process according to the feedback information. Meanwhile, the algorithm improves the sampling efficiency by mixing the samples of the current epoch with the samples of the previous epoch. Also, it captures implicit associations, especially higher-level features, and thus improves the robustness of action recognition [12].

2) MS-G3D

Liu et al. proposed a disentangled multi-scale aggregation scheme and a unified spatial temporal graph convolution (G3D) operator [33].

a) Disentangled multi-scale aggregation scheme

By defining the k-adjacency matrix, a multi-scale aggregation scheme is used to remove redundant dependencies between features of different neighborhood vertices. It addresses the problem of biased weights by eliminating the redundant dependence of closer neighborhood weighting with distant neighborhood weighting. Moreover, k-adjacency matrices are relatively sparse, which can represent the graph structure efficiently.

b) Unified spatial-temporal graph convolution (G3D) operator

A typical scenario is listed in this paper. If features are captured through independent spatial only and temporal only modules and then integrated, certain complex situations will be difficult to deal with. For example, after the propagation and aggregation of a pair of strongly correlated joints through several convolutional layers, the degree of correlation between them are reduced and the redundant information is increased. To solve this problem, a convolution operator, named G3D, is designed. It utilizes an extended sliding window to skip the smooth stream of information across space-time connections. The design of G3D refers to the definition of receptive fields in 3D convolutional network [38]. Best performance is achieved when G3D is applied to long-range and factorized modules.

Finally, feature extractor (MS-G3D) is built by fusing the above two solutions. Unlike most other approaches, this paper presents a disentanglement approach that has achieved outstanding performance. Another point is that dilated convolutions [39] are applied to multi-scale aggregation, which effectively controls the complexity of the network architecture.

Same as 2s-AGCN [24], the authors improved the efficiency of recognition by fusing joint information and bone information. The difference is that MS-G3D [33] achieves higher recognition efficiency while the total number of parameters is reduced.

3) MST-GCN

The authors proposed that based on the consideration of short-range joint dependencies and short-term trajectory, existing research do not deal with modeling the distant joints relations and long-range temporal information satisfactorily. The problems mainly include two aspects. First, they introduce additional modules or adaptively learn the relationship between vertices. Second, the higher-order polynomial of adjacency matrix is adopted. A few previous methods [32][24][40][33] only partially solve these problems.

The purpose of this model [34] is to solve the two problems mentioned above. To be specific, a multi-scale spatial graph convolution (MS-GC) module and a multi-scale temporal graph convolution (MT-GC) module are defined. Then, a multi-scale spatial temporal graph convolution network (MST-GCN) is proposed by stacking the block combined with MS-GC and MT-GC.

a) MS-GC module

The key inspiration is from Res2Net [41], which is a successful practice in CNN. In contrast to Res2Net [41], the MS-GC module also exploits smaller groups of filters to split a feature into fragments in a channel dimension. Meanwhile, spatial graph convolution is used to replace 3×3 convolution in CNN design, thus forming hierarchical residual-like architecture. Without adding more parameters, this design can capture larger receptive fields and obtain both short-range and long-range joints. Finally, all fragments will be concatenated to help model convergence.

b) MT-GC module

MT-GC module is an extension of MS-GC module in temporal. It utilizes a set of sub-temporal graph convolutions to form hierarchical residual-like structures.

c) MST-GCN

MST-GCN is a network structure formed by stacking modules composed of MS-GC and MT-GC. It is worth mentioning that another method is to capture long range features using multi-scale is MS-G3D [33]. Both of them significantly increase temporal receptive fields, but they adopt different ways. MS-G3D [33] utilizes paralleled 3×1 kernel sizes combined with dilated window, while MST-GCN [34] uses single block of a hierarchical architecture.

Additional, as an extended application of Res2Net [41] on GCN, the following two points are worth further discussion: 1) Res2Net [41] can be used as an independent block to plug into other mainstream CNN backbone such as ResNet [42] or ResNeXt [43] to expand receptive fields. 2) Res2Net [41] discusses the integration with cardinality dimension [43] and squeeze and congestion [45] blocks. However, no relevant experiments about how to adopt above two points in GCN are conducted in this paper.

C. Two-Stream GCN

1) 2s-AGCN

The authors of this paper [24] firstly put forward two deficiencies of ST-GCN [19]. First, the topology of graph is the same at each layer, which therefore lacks the flexibility. Second, compared with the data-driven graph structure, it is difficult for the fixed graph structure to get the optimal value for all the samples in various categories.

Then, several improvement strategies are creatively proposed. It is worth learning from the following two points: 1) Although ST-GCN [19] can process 2D and 3D skeleton datasets, only first-order information is considered. Apart from first-order information, in this paper, bone information including length and direction information is further explored. Therefore, an adaptive two-stream network is constructed together with joint information. 2) In view of the low flexibility of ST-GCN [19], the adjacency matrix is divided into three parts in this paper. Part 1, same as the definition of ST-GCN [19], represents the original structure of human body. Part 2, the network training completely relies on the data-driven method, which has no restrictions on capturing features of objects. Part 3, gaussian function is utilizes to compute the strength of connection between two joints. The method of data dependence is adopted, and a unique graph can be learned for each data sample.

Additionally, this output is another key research achievement after ST-GCN [19], many subsequent studies [26][30][27][33][35] are based on this paper.

2) DGNN

In this paper [35], the directed acyclic graph (DAG) is explored to represent the relationship between joints and bones. Meanwhile, directed graph neural network (DGNN) is designed to analyze and predict joints and bones and their relationship. Finally, spatial and temporal information is sent to a two-stream network for action recognition task.

The major contribution is that the skeleton-based data is defined as DAG instead of undirected graphs defined by other GCN-based methods. It takes the root vertex as the center of gravity in the human body skeleton and defines the direction of each bone accordingly. It also includes how to update the acquired features by two functions, named updating function and aggregation function.

Based on the consideration of directed graph, incidence matrix is innovatively adopted while implementing the directed graph network (DGN) block. Furthermore, the drawback of adjacency matrix used by the conventional methods [19][24] are analyzed in detail. It provides an important reference for skeleton-based human action recognition using directed graph.

It should be mentioned that its computational complexity is extremely high, with exceeding 100 giga floating-number operations (GFLOPs) [27], due to the application of directed graph and the fusion strategy of multi-streams.

3) SDGCN

In this model [26], the ideas of two well-known networks, ResNet [42] and DenseNet [46], are adopted to enhance the capability of GCN in skeleton data.

Specifically, a cross domain spatial residual layer is designed to build residual blocks. However, how to apply the key function as avoiding problems of vanishing/exploding gradients in GCN is not mentioned here. The paper does not even specify the number of layers of its network. Therefore, there may be room for further research on GCN by using residual networks.

In combination with DenseNet [46], feature-map of each layer accumulating data of all previous layers is applied to capture global information, which is undoubtedly very effective. Intuitively, this design can handle learning problems such as the relationship between distant bones and joints. However, the authors do not conduct experiments on this angle.

In addition, for the common feature of the two previous works [42][46] i.e., reducing the possibility of vanishing-gradient problem, this paper has not given a discussion.

To sum up, residual networks and dense networks are introduced into GCN and well combined in this paper. Meanwhile, there are still some open problems worthy of further discussion.

D. Attention GCN

1) STGR

Li et al. introduce a Spatio-temporal graph routing (STGR) [25], which adaptively learns the internal higher-order connections of physically separated joints.

It also aims at the problem that the fixed skeleton structure of ST-GCN [19] is not conducive to feature learning. The STGR [25] can acquire both spatial and temporal dependencies between each pair of joints by adopting two sub-networks respectively, i.e., the spatial graph router (SGR) and temporal graph router (TGR). Then, by acquiring the dynamic graph topology, its model will be composed of STGR [25] and ST-GCN [19].

In addition, this paper presents several unique observations. Fox example, the joint connections in different positions are unbalanced. When torso joints are overly smooth, limb joints may remain in a less smooth state, which is leading to difficulties in sharing features between the two limb joints. Then, while constructing of SGR, the idea of sub-group is proposed. The skeleton joints are divided into torso joints and non-torso joints, and the results of different partitioning strategies are compared and analyzed. The empirical data for capturing the optimal solution are obtained. Another point of view is that highly correlated joints tend to mean that they are also more closely related in feature learning. These fresh viewpoints are worthy of reference and discussion in future research.

However, since the human body is a whole, so some intrinsic semantic information may be lost when the human skeleton is divided into several parts.

2) STF

The authors of this paper [31] mainly point out the following two drawbacks in other methods: 1) Most existing methods do not explicitly embed higher-order spatial temporal importance into the spatial connection of vertices. 2) The advantage of using the attention mechanism in identifying action sequences is not fully utilized.

Then, improvement strategies for the above two aspects are proposed: 1) A To-a-T Spatio-temporal Focus (STF) module is proposed with a re-defined adjacency matrix which can model the higher-order spatial temporal dynamics. It also shows the importance of the input skeleton sequence. 2) STF exploration loss, STF divergence loss and STF coherence loss on the gradient based spatial temporal focus are defined. These loss terms can ensure that the prediction of classifier can be based on all the key vertices and predict various classes according to different human body parts. At the same time, among the stacked STF modules, using the focus of the high-level GCN module to help the learning of the low-level GCN module.

It is worth noting that part of this work is built on MS-G3D [33]. It points that the k-adjacency matrix defined by MS-G3D [33] cannot model the high-order relationship in the spatial temporal domain. While this model defines a dynamic adjacency matrix based on the k-adjacency matrix to solve this problem.

E. Encoder-Decoder GCN

1) AS-GCN

In AS-GCN [32], the meaning of joint information is identical to other conventional methods. However, bone information is expressed as actional links (A-links) and structural links (S-links) respectively. Where A-links are used to capture latent dependencies between any joints, and S-links are used to represent higher-order relationships of actions.

Then an actional-structural graph convolution (ASGC) is proposed. Meanwhile, temporal convolution network (TCN) [23] is used to capture spatial and temporal features respectively.

The main feature of this paper is that the encoder decoder framework of neural relational inference (NRI) model [47] is combined with the design of generation of A-links. And gated recurrent unit (GRU) [48] is utilized in the design of decoder module.

In addition, this work adds a head for prediction purpose. That is, while completing the recognition task, precise prediction can be made to the future pose.

F. Misc.

1) Shift-GCN

Cheng et al. [27] come up with two common problems in conventional GCN architectures. First, the computing complexity is generally heavy. Second, the receptive fields of spatial and temporal graph are predefined heuristically, in other words, they lack flexibility and can be optimized.

The solution to above problems is to refer to the idea of shift convolution [49] applied in CNN and adapt this practice in GCN. It is another successful example of migrating a CNN innovation to GCN domain. At the same time, lightweight point-wise convolutions are used to solve the problem of low flexibility of receptive fields.

In this paper, two solutions for building of spatial skeleton graph and temporal skeleton graph are presented respectively. The conventional scheme and the proposed method are compared and analyzed in detail, which is very helpful to understand the idea of this work.

IV. Development Framework

Based on the critical analyses carried out in this research, a skeleton and GCN-based human action recognition framework is proposed. Its operational process can be divided into four stages as shown in Fig. 2. The representative networks at each stage are ST-GCN [19], 2s-AGCN [24], MS-G3D [33] and MST-GCN [34] respectively.

A summary of each development process as following:

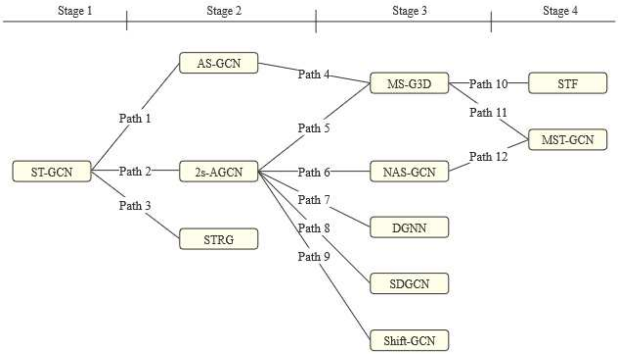

Figure 2. Development framework

A. Path 1, from ST-GCN to AS-GCN

AS-GCN [32] designs actional-links and structural-links to address incapacity of ST-GCN [19] to capture relationships between distant joints (7th drawback), obtaining action-specific latent dependencies and representing higher order relationships respectively.

B. Path 2, from ST-GCN to 2s-AGCN

2s-AGCN [24] points out four directions (2nd to 5th drawback) in which ST-GCN [19] can be improved and proposes two major improvement schemes. First, a data-driven approach is used to parameterize both the global graph and individual graph to extend the flexibility of the model. The second is to make use of the second-order information of the skeleton data and construct an adaptive graph convolutional network for prediction.

C. Path 3, from ST-GCN to STGR

STGR [25] puts forward the idea of sub-group in view of the two shortcomings (2nd and 6th drawback) of ST-GCN [19] and divided human body joints into torso joints and non-torso joints. Then, the results of different partitioning strategies are compared and analyzed, and the empirical data that can get the optimal are obtained.

D. Path 4, AS-GCN to MS-G3D

The adjacency matrix defined in AS-GCN [32] has bias in local region and vertices with high degree, it is not conducive to capture the dependencies of long-range vertices. MS-G3D [33] defines a k-adjacency matrix to address this problem, which is also more efficient.

E. Path 5, from 2s-AGCN to MS-G3D

Combining the ideas of 2s-AGCN [24] and DGNN [35], MS-G3D [33] adds a learnable graph residual mask to dynamically handle edges under the premise of the defined k-adjacency matrix. It optimizes the prediction for all possible actions and inhibits the biased weighting problem to certain extent.

F. Path 6, from 2s-AGCN to NAS-GCN

NAS-GCN [30] refers to the idea of 2s-AGCN [24], that is using both first order and second order information simultaneously and fusing the results of them to do prediction. At the same time, the improvement for two disadvantages (1st and 3rd drawback) of ST-GCN [19] is proposed.

G. Path 7, from 2s-AGCN to DGNN

DGNN [35] borrows the idea from two-stream network architecture [1][24] and makes predictions through two streams. The difference is that the method of fusing spatial stream and motion stream is used here to improve performance. In addition, a solution to the inflexibility of receptive fields in conventional GCN (9th drawback) is also provided.

H. Path 8, from 2s-AGCN to SDGCN

SDGCN [26] introduces a scheme by combining the advantages of ResNet [42] and DenseNet [46]. Taking ST-GCN [19] and 2s-AGCN [24] as baseline, the effectiveness of the method is verified by the experimental results discussion.

I. Path 9, from 2s-AGCN to Shift-GCN

Aiming at several drawbacks (7th to 9th drawback) of conventional GCN, Shift-GCN [27] is proposed to apply shift convolution [49] to GCN. In particular, the proposed non-local shift graph convolution can significantly reduce the burden of computation. It is one of the few improvement schemes for the 8th drawback. At the same time, more flexibility has been achieved in defining the receptive fields.

J. Path 10, from MS-G3D to STF

The work of STF [31] is partially built on MS-G3D [33]. The k-adjacency matrix defined in MS-G3D [33] cannot model the high-order relationship in the spatial temporal domain, then a dynamic adjacency matrix based on the k-adjacency matrix is defined to solve this problem.

K. Path 11, from MS-G3D to MST-GCN

To obtain multi-scale spatial information, earlier methods [32][24][40][33] utilize methods such as applying higher-order polynomials to the skeleton data adjacency matrix, which achieve good performance, but also greatly increase the computational complexity. As the transition of Res2Net [41] in GCN, MST-GCN [34] utilizes subnet and hierarchical residual architecture to well control the overall computation burden.

Meanwhile, in terms of multi-scale temporal modeling, MST-GCN [34] adopts a single block, which is different from the previous method [33], by using 3×1 kernel size. And the accumulation of short- and long-range information is carried out through hierarchical architecture.

L. Path 12, from NAS-GCN to MST-GCN

While certain skeleton data-based methods [30][33] generate multi-scale structural features through higher-order polynomials, MST-GCN [34] transfers the idea of Res2Net [41] from CNN to GCN. It uses residual connections to stack several sub-graph convolutions to capture short range vertices dependencies and distant vertices relations simultaneously.

In addition, in the process of temporal modeling, compared with the earlier methods [19][32][24][30], which used fixed kernel size, MST-GCN [34] applies sub-temporal graph convolutions. This approach can increase the temporal receptive fields, and finally assign the model with multi-scale representation capability of temporal information.

V. Performance Comparison

A. Datasets

The most used datasets for human action recognition with skeleton-based data are Kinetics-Skeleton [28], NTU RGB+D [29], and NTU RGB+D 120 [50].

1) Kinetics-Skeleton

Kinetics 400 human action dataset [28] contains around 300000 video clips in 400 classes, which can be retrieved from YouTube. There are 240436 and 19794 samples for training and testing respectively. The skeleton version is converted by publicly available OpenPose [7] toolbox. Top-1 and Top-5 accuracies are reported following the conventional protocols [28][19].

2) NTU RGB+D

NTU RGB+D [29] contains 56880 action clips in 60 classes. The clips are all captured from 40 subjects with 3 camera views recorded simultaneously. The recommendation for reporting the accuracy by two ways: Cross-Subject (X-Sub) and Cross-View (X-View). For the former setting, half of the subjects are split for training and others for testing respectively; for the latter, samples captured by camera 2 and camera 3 are for training and others for testing accordingly.

3) NTU RGB+D 120

NTU RGB+D 120 [50] is an expansion of NTU RGB+D [29] in the number of subjects and action classes. It contains 114480 action clips in 120 classes, which are captured from 106 subjects with three camera views. Similarly, Cross-Subject (X-Sub) and Cross-Setup (X-Setup) are recommended as the evaluation protocol. The split principle is same with NTU RGB+D [29] for the former setting; for the latter, half out of the 32 setups are split for training and others for testing respectively.

The reason for choosing above three datasets is because the scenarios they covered are complementary. The Kinetics-Skeleton [28][7] obtained through the public available toolbox is 2D skeleton data, compared to the 3D data of NTU series since it is obtained through depth sensors. However, since the NTU series is captured in a lab environment, it has constraint data only. As per Kinetics-Skeleton, which is probably a mix of constraint and unconstraint data, as it is all from the Internet. Additionally, the Kinetics-Skeleton [28][7] contains 18 joints per human body, while 25 joints in the NTU series.

B. The Comparative Analyses

Most of subsequent studies of ST-GCN [ST-GCN] follow its experimental settings. The performance comparison of several state-of-the-art methods is shown in Table II, Table III and Table IV.

Methods typically achieve higher performance on Kinetics-Skeleton [28][7] also produce consistent results on NTU RGB+D [29]. The exception is the rankings of DGNN [35], NAS-GCN [30] and SDGCN [26].

Only a few methods with outstanding performance [27][33][34][31] on NTU RGB+D [29] gave results on

TABLE I. IMPROVEMENT SCHEMES FOR ST-GCN

	Spatial Temporal GCN	Two-Stream GCN	Attention GCN	Encoder-Decoder GCN	Misc.
1	[30]				
2		[24]	[26]		
3	[30]	[24]	[31]		
4		[24]			
5		[24]			
6			[26]		
7	[33], [34]	[26]		[32]	[27]
8					[27]
9	[33], [34]	[35]			[27]

The first column indicates the drawback ID in Section III

TABLE II. COMPARISONS OF THE TOP-1 AND TOP-5 ACCURACY ON THE KINETICS-SKELETON DATASET

Methods	Top-1 (%)	Top-5 (%)
ST-GCN [19]	30.7	52.8
STGR [25]	33.6	56.1
AS-GCN [32]	34.8	56.6
2s-AGCN [24]	36.1	58.7
DGNN [35]	36.9	59.6
NAS-GCN [30]	37.1	60.1
SDGCN [26]	37.4	60.3
MS-AAGCN [51]	37.8	61.0
MS-G3D [33]	38.0	60.9
MST-GCN [34]	38.1	60.8
STF [31]	39.9	/

TABLE III. COMPARISONS OF THE TOP-1 ACCURACY ON THE NTU RGB+D DATASET

Methods	X-View (%)	X-Sub (%)
ST-GCN [19]	88.3	81.5
STGR [25]	92.3	86.9
AS-GCN [32]	94.2	86.8
AGC-LSTM [44]	95.0	89.2
2s-AGCN [24]	95.1	88.5
SDGCN [26]	95.7	89.6
NAS-GCN [30]	95.7	89.4
DGNN [35]	96.1	89.9
MS-AAGCN [51]	96.2	90.0
MS-G3D [33]	96.2	91.5
Shift-GCN [27]	96.5	90.7
MST-GCN [34]	96.6	91.5
STF [31]	96.9	92.5

TABLE IV. COMPARISONS OF THE TOP-1 ACCURACY ON THE NTU RGB+D 120 DATASET

Methods	X-Setup (%)	X-Sub (%)
Shift-GCN [27]	87.6	85.9
MS-G3D [33]	88.4	86.9
MST-GCN [34]	88.8	87.5
STF [31]	89.9	88.9

NTU RGB+D 120 [50]. However, the overall recognition efficiency is less than 90%, indicating that there is still a lot of room for improvement.

Different from the high performance on NTU series, the efficiencies on Kinetics-Skeleton [28][7] are generally under-expected, even Top-5 achieved only around 60%. How to improve the efficiency on video-based datasets may have greater significance for end-to-end applications in the real world.

VI. CONCLUSIONS AND FUTURE WORK

This research investigated recent advancements in GCN-based human action recognition using skeletal features. It is observed that the adjacency matrix is best suited for expressing the structure of undirected graphs. One exception is DGNN [35], where the structure of a graph is represented by incidence matrix as a directed acyclic graph.

MST-GCN [34] exposes the weakness with higher-order polynomials of the skeleton adjacency matrix that generates huge number of parameters. A MS-GC module for splitting subsets is a viable way forward. The STF module [31] implements a dynamic adjacency matrix scheme, which is based on a thorough analysis of all previous adjacency matrix designs. It achieved the best performance identified in this study.

For the future direction of human action recognition research, the existing studies have laid a solid foundation. In addition, it is envisaged a rewarding arena in combining skeleton data with other data modalities.

REFERENCES

[1] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." Advances in neural information processing systems 27 (2014). pp. 568-576

[2] Buch, Shyamal, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. "Sst: Single-stream temporal action proposals." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 2911-2920. 2017.

[3] Ke, Qiuhong, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. "A new representation of skeleton sequences for 3d action recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3288-3297. 2017.

[4] Yang, Zhengyuan, Yuncheng Li, Jianchao Yang, and Jiebo Luo. "Action recognition with visual attention on skeleton images." In 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3309-3314. IEEE, 2018.

[5] Varol, Gül, Ivan Laptev, and Cordelia Schmid. "Long-term temporal convolutions for action recognition." IEEE transactions on pattern analysis and machine intelligence 40, no. 6 (2017): 1510-1517.

[6] Sun, Zehua, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. "Human action recognition from various data modalities: A review." arXiv preprint arXiv:2012.11866 (2020).

[7] Cao, Zhe, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "Realtime multi-person 2d pose estimation using part affinity fields." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7291-7299. 2017.

[8] Shotton, Jamie, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. "Real-time human pose recognition in parts from single depth images." In CVPR 2011, pp. 1297-1304. IEEE, 2011.

[9] Ahmad, Tasweer, Lianwen Jin, Xin Zhang, Songxuan Lai, Guozhi Tang, and Luojun Lin. "Graph Convolutional Neural Network for Human Action Recognition: A Comprehensive Survey." IEEE Transactions on Artificial Intelligence 2, no. 2 (2021): 128-145.

[10] Ren, Bin, Mengyuan Liu, Runwei Ding, and Hong Liu. "A survey on 3d skeleton-based action recognition using learning method." arXiv preprint arXiv:2002.05907 (2020).

[11] Xing, Yuling, and Jia Zhu. "Deep learning - based action recognition with 3D skeleton: A survey." (2021): 80-92.

[12] Feng, Liqi, Yaqin Zhao, Wenxuan Zhao, and Jiaxi Tang. "A comparative review of graph convolutional networks for human skeleton-based action recognition." Artificial Intelligence Review (2021): 1-31.

[13] Sarkar, Arya, Avinandan Banerjee, Pawan Kumar Singh, and Ram Sarkar. "3D Human Action Recognition: Through the eyes of researchers." Expert Systems with Applications (2022): 116424.

[14] Du, Yong, Wei Wang, and Liang Wang. "Hierarchical recurrent neural network for skeleton based action recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1110-1118. 2015.

[15] Liu, Jun, Amir Shahroudy, Dong Xu, and Gang Wang. "Spatio-temporal lstm with trust gates for 3d human action recognition." In European conference on computer vision, pp. 816-833. Springer, Cham, 2016.

[16] Zhang, Pengfei, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. "View adaptive recurrent neural networks for high performance human action recognition from skeleton data." In Proceedings of the IEEE international conference on computer vision, pp. 2117-2126. 2017.

[17] Sainath, Tara N., Oriol Vinyals, Andrew Senior, and Haşim Sak. "Convolutional, long short-term memory, fully connected deep neural networks." In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4580-4584. IEEE, 2015.

[18] Li, Chao, Qiaoyong Zhong, Di Xie, and Shiliang Pu. "Skeleton-based action recognition with convolutional neural networks." In 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 597-600. IEEE, 2017.

[19] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." In Thirty-second AAAI conference on artificial intelligence. 2018.

[20] Wang, Pei, Chunfeng Yuan, Weiming Hu, Bing Li, and Yanning Zhang. "Graph based skeleton motion representation and similarity measurement for action recognition." In European conference on computer vision, pp. 370-385. Springer, Cham, 2016.

[21] Dai, Jifeng, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. "Deformable convolutional networks." In Proceedings of the IEEE international conference on computer vision, pp. 764-773. 2017.

[22] Dai, Chuan, Yajuan Wei, Zhijie Xu, Minsi Chen, Ying Liu, and Jiulun Fan. "An Investigation into Performance Factors of Two-Stream I3D Networks." In 2021 26th International Conference on Automation and Computing (ICAC), pp. 211-216. IEEE, 2021.

[23] Soo Kim, Tae, and Austin Reiter. "Interpretable 3d human action analysis with temporal convolutional networks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20-28. 2017.

[24] Shi, Lei, Yifan Zhang, Jian Cheng, and Hanqing Lu. "Two-stream adaptive graph convolutional networks for skeleton-based action recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12026-12035. 2019.

[25] Li, Bin, Xi Li, Zhongfei Zhang, and Fei Wu. "Spatio-temporal graph routing for skeleton-based action recognition." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 8561-8568. 2019.

[26] Wu, Cong, Xiao-Jun Wu, and Josef Kittler. "Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition." In proceedings of the IEEE/CVF international conference on computer vision workshops, pp. 1740-1748. 2019.

[27] Cheng, Ke, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. "Skeleton-based action recognition with shift graph convolutional network." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 183-192. 2020.

[28] Kay, Will, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola et al. "The kinetics human action video dataset." arXiv preprint arXiv:1705.06950 (2017).

[29] Shahroudy, Amir, Jun Liu, Tian-Tsong Ng, and Gang Wang. "Ntu rgb+ d: A large scale dataset for 3d human activity analysis." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1010-1019. 2016.

[30] Peng, Wei, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. "Learning graph convolutional network for skeleton-based human action recognition by neural searching." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 03, pp. 2669-2676. 2020.

[31] Ke, Lipeng, Kuan-Chuan Peng, and Siwei Lyu. "Towards To-aT Spatio-Temporal Focus for Skeleton-Based Action Recognition." arXiv preprint arXiv:2202.02314 (2022).

[32] Li, Maosen, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. "Actional-structural graph convolutional networks for skeleton-based action recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3595-3603. 2019.

[33] Liu, Ziyu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. "Disentangling and unifying graph convolutions for skeleton-based action recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 143-152. 2020.

[34] Chen, Zhan, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 2, pp. 1113-1122. 2021.

[35] Shi, Lei, Yifan Zhang, Jian Cheng, and Hanqing Lu. "Skeleton-based action recognition with directed graph neural networks." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7912-7921. 2019.

[36] Zoph, Barret, and Quoc V. Le. "Neural architecture search with reinforcement learning." arXiv preprint arXiv:1611.01578 (2016).

[37] Larrañaga, Pedro, and Jose A. Lozano, eds. Estimation of distribution algorithms: A new tool for evolutionary computation. Vol. 2. Springer Science & Business Media, 2001.

[38] Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks." In Proceedings of the IEEE international conference on computer vision, pp. 4489-4497. 2015.

[39] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122 (2015).

[40] Zhang, Xikun, Chang Xu, and Dacheng Tao. "Context aware graph convolution for skeleton-based action recognition." In Proceedings of the IEEE/CVF conference

on computer vision and pattern recognition, pp. 14333-14342. 2020.

[41] Gao, Shang-Hua, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. "Res2net: A new multi-scale backbone architecture." IEEE transactions on pattern analysis and machine intelligence 43, no. 2 (2019): 652-662.

[42] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[43] Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated residual transformations for deep neural networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492-1500. 2017.

[44] Si, Chenyang, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. "An attention enhanced graph convolutional lstm network for skeleton-based action recognition." In proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1227-1236. 2019.

[45] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141. 2018.

[46] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. 2017.

[47] Kipf, Thomas, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. "Neural relational inference for interacting systems." In International Conference on Machine Learning, pp. 2688-2697. PMLR, 2018.

[48] Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[49] Wu, Bichen, Alvin Wan, Xiangyu Yue, Peter Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph Gonzalez, and Kurt Keutzer. "Shift: A zero flop, zero parameter alternative to spatial convolutions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9127-9135. 2018.

[50] Liu, Jun, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding." IEEE transactions on pattern analysis and machine intelligence 42, no. 10 (2019): 2684-2701.

[51] Shi, Lei, Yifan Zhang, Jian Cheng, and Hanqing Lu. "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks." IEEE Transactions on Image Processing 29 (2020): 9532-9545.