





Graph Convolutional Neural Network for Human Action Recognition: A Comprehensive Survey

Tasweer Ahmad , Lianwen Jin , *Member, IEEE*, Xin Zhang , *Member, IEEE*, Songxuan Lai , Guozhi Tang, and Luojun Lin

Abstract—Video-based human action recognition is one of the most important and challenging areas of research in the field of computer vision. Human action recognition has found many pragmatic applications in video surveillance, human–computer interaction, entertainment, autonomous driving, etc. Owing to the recent development of deep learning methods for human action recognition, the performance of action recognition has significantly enhanced for challenging datasets. Deep learning techniques are mainly used for recognizing actions in images and videos comprising of Euclidean data. A recent development in deep learning methods is the extension of these techniques to non-Euclidean data or graph data with many nodes and edges. Human body skeleton resembles a graph, therefore, the graph convolutional network (GCN) is applicable to the non-Euclidean body skeleton. In the past few years, GCN has emerged as an important tool for skeleton-based action recognition. Therefore, we conduct a survey using GCN methods for action recognition. Herein, we present a comprehensive overview of recent GCN techniques for action recognition, propose a taxonomy for the categorization of GCN techniques for action recognition, carry out a detailed study of the benchmark datasets, enlist relevant resources and open-source codes, and finally provide an outline for future research directions and trends. To the best of authors’ knowledge, this is the first survey for action recognition using GCN techniques.

Impact Statement—Graph convolutional neural networks have marked a great progress in recent years. There is a similarity between body skeleton and a graph; therefore, GCNs have been widely used for skeleton-based action recognition. In this article, we summarize recent graph-based action recognition techniques, provide a deeper insight of these methods, list source-codes and

available resources. This article will help the researchers to develop a basic understanding of graph convolutional methods for action recognition, benefit from useful resources, and think about future directions.

Index Terms—Deep learning techniques, graph convolution network, human–robot interaction.

I. INTRODUCTION

HUMAN action recognition has been an important research topic over the past decades, and many new methods have been proposed for action recognition. The task of action recognition is challenging owing to background clutter, illumination variation, scale-variation, view-point variation, occlusion, and variation in the motion speed and so on. With the advent of deep learning techniques, most of these issues have been well addressed to some extent and significant progress has been achieved in human action recognition; however, the problems in action recognition have not been fully solved. With the advancements in human action recognition approaches, benchmark datasets have become larger and more challenging over the years.

For the task of action recognition, various modalities have been investigated, e.g., raw RGB images [1]–[3], skeleton-modality [4]–[6], and optical-flow modalities [2], [7]–[9], etc. The skeleton modality yielded a significant success, and numerous skeleton-based action recognition approaches have recently been proposed [10]–[12]. It has gained significance over other modalities because it contains the following:

- 1) spatial information, which entails a strong correlation between the joint node and its adjacent nodes for modeling intrinsic structural information in a frame;
- 2) temporal information, which models strong temporal correlation within the frames in an intraframe manner, additionally;
- 3) the co-occurrence relationship is elegantly modeled using the skeleton modality when the spatial and temporal domains are simultaneously taken into account.

In recent, convolutional neural networks have been applied to Euclidean data for object recognition [13], [14] and action recognition, [2], [4], [9]. In line with this study, convolutional neural networks have been extended to non-Euclidean data such as social network data [15], [16], [17] gene and protein sequence data [18], [19] and e-commerce data (modeling the relationship

Manuscript received September 21, 2020; revised January 11, 2021 and March 4, 2021; accepted April 22, 2021. Date of publication April 30, 2021; date of current version August 20, 2021. This work was supported in part by the National Natural Science Foundation China (NSFC Grants 61936003 and 61771199), in part by the Natural Science Foundation of Guangdong Province (GD-NSF Nos. 2017A030312006 and 2018A030313295), in part by the Fundamental Research Funds for the Central Universities (x2dxD2190570), and in part by the Science and Technology Program of Guangzhou (2018-1002-SF-0561). This article was recommended for publication by Associate Editor D. Zhao upon evaluation of the reviewers’ comments. (*Corresponding author: Lianwen Jin.*)

Tasweer Ahmad is with the School of Electronics, and Information Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Electrical and Computer Engineering Department, COMSATS University Islamabad, Islamabad 45550, Pakistan (e-mail: tasveerahmad@gmail.com).

Lianwen Jin is with the School of Electronics, and Information Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Guangdong Artificial Intelligence and Digital Economy Laboratory, Guangzhou 510335, China (e-mail: lianwen.jin@gmail.com).

Xin Zhang, Songxuan Lai, and Guozhi Tang are with the School of Electronics, and Information Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: eexinzhang@scut.edu.cn; linluojun2009@126.com; gztang@126.com).

Luojun Lin is with the School of Mathematics and Computer Science, Fuzhou University, Fuzhou 350122, China (e-mail: eesxlai@qq.com).

Digital Object Identifier 10.1109/TAI.2021.3076974

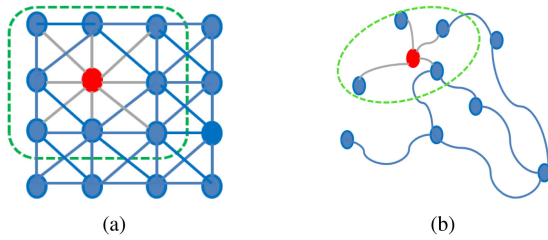


Fig. 1. Illustration of 2-D (Euclidean) convolution versus Graph convolution [26]. (a) Convolution operation in euclidean domain. (b) Convolution operation on a graph.

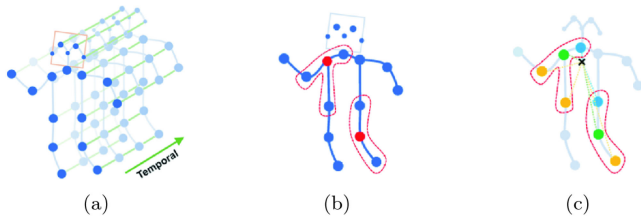


Fig. 2. Illustration of spatio-temporal Stacking of skeleton frames [22]. (a) Spatio-temporal graph for a skeleton-sequence. (b) Distance partitioning, root node is at distance 0, while neighboring nodes are at distance 1. (c) Spatial partitioning, nodes are labeled according to their distances to the skeleton gravity center.

between products and customer behaviors) [20]. The convolutional neural networks are also extended to non-Euclidean space using the graph convolutional network (GCN), as proposed in [21]. An analogy between the convolutional neural network and GCN is illustrated in Fig. 1. By applying this idea of GCN to non-Euclidean data, researchers have investigated GCN for skeleton-based action recognition. GCN-based action recognition involves GCN for recognizing different actions, and body skeleton is the input modality for such kind of action recognition architectures. In the pioneering work for modeling skeleton-based action recognition using GCN [22], body skeletons are represented as spatio-temporal graphs over the length of a video sequence. This concept is illustrated in Fig. 2. Over the past three years, numerous approaches have evolved for skeleton-based action recognition using GCNs.

Accordingly, there have been many surveys about the development of GCN. The first review of GCN by Bronstein *et al.* [23], mainly concerned about convolution operation on graphs, as well as its significance and applications. Hamilton *et al.* [24] investigated representation learning and network embedding on graphs. Lee *et al.* [25] conducted a comprehensive survey on graph attention networks. A survey with new taxonomy for GCN was proposed by Wu *et al.* [26]. Zhang *et al.* [27] performed a comprehensive survey of recently developed GCN techniques. Zhang *et al.* [28] categorized graph convolution methods into five subclasses graph recurrent neural networks (RNNs), GCNs, graph autoencoders, graph reinforcement learning, and graph adversarial methods. This abovementioned literature highlighted the recent developments in general GCN architectures and advocated its applications in different areas.

Furthermore, a 3-D Kinect sensor-based human action recognition techniques are examined in a survey by Wang *et al.* [29]. This survey focused on convolutional neural networks, recurrent neural networks (RNNs), and GCN techniques, which have significantly contributed to 3-D Kinect sensor-based action recognition. Wang *et al.* [29] compares the performance of only ten recent state-of-the-art action recognition methods. Moreover, they emphasize on a comparison using hand-crafted features versus deep learning features. However, skeleton-based action recognition using GCN is not the main focus of their paper. Meanwhile, this study is more adherent and investigative to skeleton-based action recognition using graph convolution in terms of categorization, mathematical modeling, performance comparison, and future guidelines. In another 3-D human action recognition survey [30], skeleton-based action recognition techniques are also divided into the following three main categories: convolutional neural networks, RNNs, and GCNs. However, only few GCN-based methods are mentioned in this article. Moreover, the work in [30] only highlights NTU-RGBD-60 and NTU-RGBD-120 datasets and all performance comparisons for different methods are made on these two datasets. Different from all existing surveys, this article specifically covers skeleton-based action recognition techniques using graph convolution and addresses the development and challenges for these techniques. We also provide substantial description for the performance comparison of different skeleton datasets.

In this article, we summarize up-to-date research on human action recognition using GCN. For the literature survey and data collection, it has been strictly adherent that

- 1) the article is related to GCN-based action recognition;
- 2) manuscript is published in well-reputed journal or conference;
- 3) the proposed methods are properly presented with experimentation and explanation of results.

Research Contribution. The main contribution of this article to the literature is summarized as follows.

- 1) *Comprehensive Overview:* We present a comprehensive overview of the contemporary techniques for GCN-based action recognition. We provide insightful discussions, mathematical models and highlight their contributions and limitations.
- 2) *Proposed Taxonomy:* We categorize GCN-based action recognition into the following five groups:
 - 1) spatio-temporal GCN;
 - 2) RNN-attention GCN;
 - 3) two-multistream GCN;
 - 4) encoder-decoder GCN;
 - 5) miscellaneous GCN architectures.

To the best of authors' knowledge, we are the first to develop such taxonomy for GCN-based action recognition and report a survey for action recognition.

- 3) *Useful Resources:* We list useful resources such as benchmark datasets, GCN libraries, open-sources codes, and online trained models weights, etc.
- 4) *Future Directions:* Finally, we identify the limitations of existing models and briefly demonstrate the possible

future directions and guidelines for GCN-based action recognition.

The remaining article is organized as follows. A background study and some basic terms are defined in Section II. In Section III, we present the taxonomy of GCN-based action recognition techniques. We also explain different methods in each category. Section IV presents skeletal benchmark datasets for performance evaluation using various GCN techniques. The resources, source codes, and tutorial guidelines are presented in Section V. Finally, Section VI concludes this article with recent trends and future directions.

II. BACKGROUND STUDY AND DEFINITIONS

In this section, first we come up with preliminary background study on human action recognition and then we define some basic terms for this article.

A. Background Study

In the literature, action recognition approaches are summarized as RGB and depth techniques, deep learning-based representation methods, human-human, and human-object interaction frameworks and recent graph neural network architectures [31]. Initial action recognition work included RGB-images for spatial features and hand-crafted temporal features spatial temporal interest points (STIPs), scale-invariant feature transform (SIFT) for action recognition [32] [33]. With the strong representational power of deep learning-based methods, RGB and depth data are incorporated for action recognition. Deep learning approaches modeled spatial features from raw RGB images, while temporal features are modeled using optical-flow or recurrent architecture [CNN+long short-term memory (LSTM)]. Using body skeleton, CNN-based action recognition faces some serious problems, such as size and speed of model, occlusion, and view-point variation [34], [35]. So, there is on going research in order to solve these challenges. Vision-based action recognition can also take advantages from scene and object knowledge. Different methods have made attempt to recognize human actions using human-human interaction and human-object interaction [36]–[38]. Such kind of frameworks require scene object detection and interaction models [39]. Since body skeleton resembles to a topological graph and there is remarkable recent development in graph convolution; therefore, the task of skeleton-based action recognition has also been addressed using GCN. GCN treats body skeleton as a graph, instead of sequence vector or matrix image. Recently a great number of excellent GCN-based action recognition techniques have been proposed. However, such approaches face a serious problem that how to represent a raw skeleton data into a graph [30].

如何在图中表示人体骨架数据

B. Spectral Versus Spatial GCN

Spectral graph convolution is implemented by the multiplication of two signals, which in turn requires Chebyshev polynomials for efficient implementation [21]. Chebyshev polynomials are studied as orthogonal polynomials, which are very good at approximating functions. Referring to [40], spectral domain

用来逼近函数

GCN involves eigen decomposition of graph Laplacian matrix, $\nabla = D^{-1/2}AD^{-1/2}$. Mathematically, spectral domain GCN is defined as

$$z = g(\theta') \star x = \sum_{m=0}^M \theta'_m T_m(\tilde{\nabla})x \quad (1)$$

for $\tilde{\nabla} = 2/\lambda \times \nabla - I_{N \times N}$, where λ is a vector of Chebyshev coefficients up to M th order. Since the abovementioned expression is an M th order Laplacian polynomial, which actually depends on the nodes that are at maximum M steps away from the central node (M th order neighborhood). ChebNet [41] and, CayleyNet [42], [43] provide a detailed discussion on spectral graph convolution.

空间的图卷积聚焦于顶点的局部邻居

Spatial Graph convolution is implemented on local neighborhood of nodes and understands the properties of local neighboring nodes. Spectral convolution takes a lot of computational time, which could be reduced by using spatial GCN. Spatial GCN is a better generalization of spectral convolution and produces state-of-the-art results for numerous applications. Referring to (1), we formulate the spatial GCN by limiting the layer-wise convolution operation to $M = 1$, which leads to a linear convolution. Intuitively speaking, it computes the convolution operation among the neighboring nodes one-step from the central nodes. For $M = 1$ and $\lambda = 2$, (1) can be rewritten as a linear convolution of adjacent neighboring nodes

$$z = g(\theta) \star x = \theta(I_{N \times N} + D^{-1/2}AD^{-1/2})x. \quad (2)$$

Further normalized formulation of the abovementioned expression can be rewritten as

$$Z = \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} X \Theta \quad (3)$$

where $\Theta \in \mathbb{R}^{C \times F'}$ is a matrix of filter parameters and, $Z \in \mathbb{R}^{N \times F'}$ is a convolved signal matrix. In addition, C is number of input channels, and F is the feature dimension. Equation (3) contains one-hop neighbor information for graph convolution. For $M = 2$ in (1), it could incorporate second-hop neighbor information, i.e., neighbors of neighbors information. GraphSage [16], and MoNet [44], [43] are good examples of spatial graph convolution and provide deeper insights.

C. Definitions

In this section, we define some basic terms, notations, and mathematical symbols that we will use during the discussions. A detailed list of notations and symbols used in this article are provided in Table I. In the literature, some of authors have used different notations to denote the same parameters, but in this article, we represent all those parameters with generalized notations. For example, some authors use D to denote degree matrix of a graph, while others used symbol, λ ; therefore, we have tried to keep the notations generalized and simple, therefore, use the notation D to represent degree matrix in this article.

Definition 1: In this article, we represent a graph as $G = (V, E)$, where V is a set of nodes and E is a set of edges. We define an edge $e_{ij} = (v_i, v_j)$ from node v_i to v_j . Moreover, we

切比雪夫多项式

TABLE I
NOTATION AND SYMBOLS

Notations	Description
A	Adjacency matrix
b	weight bias
D	Degree matrix
E	Set of Edges in a Graph
G	A Graph
H^l	Hidden layer l
M	number of persons in a frame
T	number of frames
U	Fourier Basis
v	Set of nodes/vertices
W^l	Weight-matrix for layer l
X_{in}	Input feature layer X
X_e	Input as edges
X_v	Input as node/vertex
Y_{out}	Output Y
d	Input feature dimension
σ	Sigmoid activation function
\circ	Hadamard Product
λ_i	i -th Eigen value
L	Loss function
\otimes	Element-wise Product
∇	Graph Laplacian

define the adjacency-matrix A as an $N \times N$ matrix such that

$$A_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in S \\ 0 & \text{if } e_{ij} \notin S. \end{cases} \quad (4)$$

Definition 2: For this article, we explicitly define the skeleton-based action recognition, where the input to GCN is a 2-D or 3-D body skeleton. Body-skeleton in each frame, t , is represented as a graph $G_t = (V, E)_t$.

III. NEW TAXONOMY FOR GCN-BASED ACTION RECOGNITION

We propose a new taxonomy for the categorization of different action recognition methods using GCN. We broadly classify GCN-based action recognition tasks into the following five main categories:

- 1) spatio-temporal GCN;
- 2) recurrent-attention GCN;
- 3) two-multistream GCN;
- 4) encoder-decoder GCN;
- 5) miscellaneous GCN.

We provide a schematic diagram in Fig. 3, which explains the basic architecture of these categories.

Spatio-temporal GCN is the pioneering work for skeleton-based action recognition. In this method, human body skeleton is treated as a fixed graph upto N -frames. Spatial relationship between joints is established by spatial edges within frame, whilst temporal relationship between joints is established by temporal edges between frames.

Recurrent-attention GCN includes the recurrent network or attention network in the architecture. Attention networks are used to emphasize salient portion of a body skeleton, which are mostly implemented by using recurrent or LSTM architecture.

Two-multistream GCN attempts to model complementary information (of joints and bones) of body skeleton for action

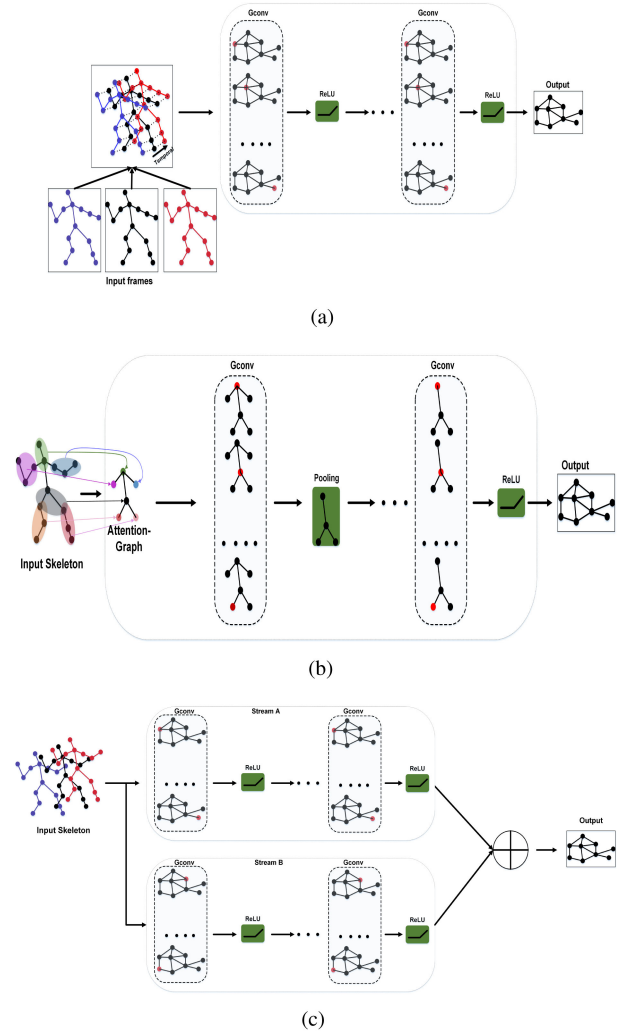


Fig. 3. Schematic diagram for illustration of different categories of GCN for action recognition. (a) Schematic for spatio-temporal GCN. (b) Schematic for recurrent-attention GCN. (c) Schematic for 2-stream GCN.

recognition. Two-streams or multistreams architectures comprise of separate streams of graph convolutions, which are later concatenated and softmax function is used for classification.

Encoder-decoder GCN are unsupervised learning methods for encoding nodes/graphs into a latent vector space (learning a network embedding) and then graphs are reconstructed from the encoded information.

A. Spatio-Temporal Graph Convolutional Network (ST-GCN)

ST-GCN takes the input body skeleton as a spatio-temporal graph having spatial and temporal edges between body skeleton. A spatio-temporal graph is represented as, $\zeta = \{G_1, G_2, \dots, G_N\}$, where G_1, G_2, \dots, G_N is the input skeleton-graph in each frame and N is the total number of frames. These spatio-temporal graphs are excited to GCN. A schematic diagram for this method is shown in Fig. 3(a). The main advantage of ST-GCN is that it has generic design for automatically learning both spatial and temporal patterns from the input data.

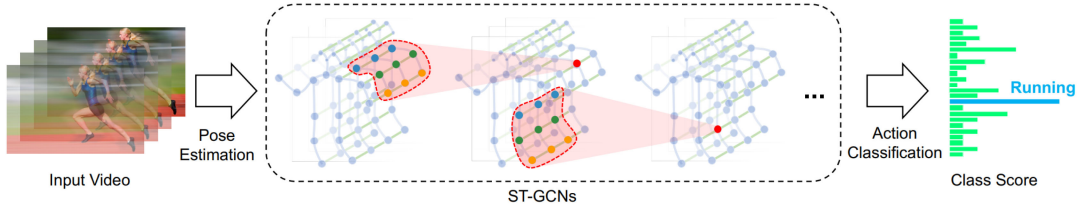


Fig. 4. Illustration of spatio-temporal GCN [22]. Multiple layers of ST-GCN are applied, which gradually generate higher level feature maps on the graph. Finally, standard softmax classifier is used for classification to the corresponding action category.

Second, spatio-temporal GCN has greater expressive power and generalization capability; moreover, it works in a complementary fashion, analogous to RGB, and optical flow information. The pioneering work using ST-GCN was proposed by Yan *et al.* [22]. The authors used an input video skeleton having spatio-temporal graph with many nodes and edges. Within-frame edges contribute to spatial details whereas between-frames edges provide the temporal information. Such spatio-temporal representation is intuitive for modeling skeleton-based action recognition, as shown in Fig. 4. The following mathematical model was devised for this spatio-temporal architecture:

$$Y_{\text{out}} = \sum_j D^{-1/2} A_j D^{-1/2} X_{\text{in}} W_j^l. \quad (5)$$

ST-GCN draws the intuition from 2-D convolution for the implementation of sampling function and weighting function. The feature map in CNN is usually of a reduced size, but for GCN, appropriate padding and setting stride as 1 results the output feature map of the same size as that of the input feature map. In fact, ST-GCN is a powerful idea that provides the foundation for STGR [45], part-based GCN (PB-GCN) [46], and ST-Graph sparsification [47].

Chaolong *et al.* [48] extended this idea of spatio-temporal GCN for skeleton-based action recognition by combining the benefits of local convolutional filtering and the sequence learning ability of autoregressive moving average. This multiscale local GCN consists of local receptive fields and recursively performs graph convolution in spatial and temporal domains. In skeleton-based action recognition, distant joints contain important information for action recognition; for example, in clapping action, there is an important relationship between distant joints of two hands. Li *et al.* [45] addressed this problem by devising a spatio-temporal graph routing (STGR) methodology, which can adaptively learn higher order relationships between distant skeleton joints. STGR comprises of spatial graph routing (SGR) and temporal graph routing (TGR). SGR includes subgroup clusters for exploiting the connectivity relationship among distant joints and TGR emphasizes on the structural information by correlating the temporal joint trajectories. STGR yields a well-balanced receptive field for different joints and is effective for skeleton-based action recognition, defined as

$$\{g^{\text{spat}}, g^{\text{temp}}\} = f_{\text{STGR}}(X_{\text{in}}; \theta^{\text{spat}}, \theta^{\text{temp}}) \quad (6)$$

where g^{spat} and g^{temp} are spatial and temporal graph topologies, respectively. θ^{spat} and θ^{temp} represent the corresponding parameters. g^{spat} and g^{temp} are concatenated with the g^{default} to form a graph set $S = \{g^{\text{default}}, g^{\text{spat}}, g^{\text{temp}}\}$ for GCN. This concept of STGR is vividly explained in Fig. 5. **Multiscale and Body-Parts Feature aggregation:** A technique combining the multiscale features from all levels of skeleton feature hierarchy was proposed by Parsa *et al.* [49]. These multiscale features work in a pyramid fashion, where features at different hierarchical levels have different semantic meaning. Such pyramid of features can capture the correlation between body parts and limbs. This method involves ergonomic risk index and can work for real-time action recognition. In ST-GCN, the importance of different body parts was model by using PB-GCN in [46]. It is argued that human skeletons can be represented as sparse graphs with nodes and edges; therefore, inspired by deformable part-based models, the PB-GCN can be applied for skeleton-based action recognition. A novel ST-GCN method using graph representation learning has been proposed by [50]. This proposed representation learning enhances the relationship between limbs and the center of gravity of body-skeleton for action recognition. The proposed framework is excited with RGB-images and 3D-body skeleton. The proposed method is evaluated on benchmark action recognition dataset and it achieved promising results. Peng *et al.* [51] discovered neural architecture search (NAS) with an automatically designed GCN for ST-GCN action recognition. NAS replaces the fixed graph with a dynamic graph and better models skeleton-based action recognition. This technology includes spatial-temporal block that generates dynamic graphs using node correlation of different function blocks. Moreover, this technique makes the GCN receptive field larger by computing a higher order Chebyshev polynomial approximation. The proposed method is evaluated on large-scale benchmark dataset and this technique performs better for a significant margin.

Wu *et al.* [52] devised a new idea of spatial residual layer for capturing the rich spatio-temporal information along with a dense connection block. The spatial residual layer is used to extract precise and accurate spatio-temporal information, thus, minimizes the training complexity of feature fusion. This architecture integrates 2-D spatial convolution with 1-D temporal convolution for spatio-temporal feature representation for a skeleton-based action recognition. In existing methods, body-parts features are modeled by merely joints position and, thus, inherent structured correlations among the body-parts are not

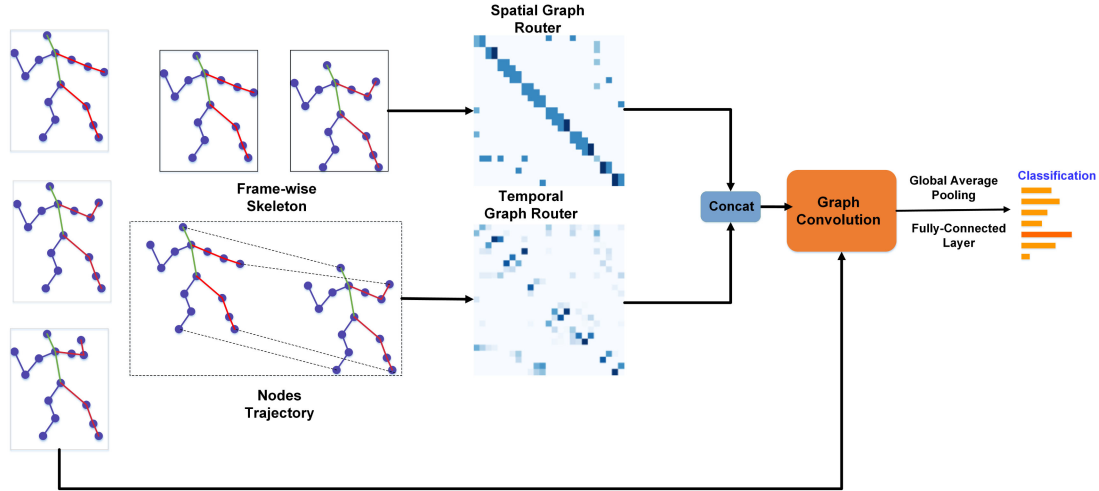


Fig. 5. Illustration of STGR [45]. spatial Graph Router (SGR) and temporal Graph Router (TGR) produce new skeleton-joint-connectivity graphs, respectively. ST-GCN receives these graphs and outputs action class.

fully emphasized. Therefore, in order to model these problems, Shi *et al.* [53] devised pose-based GCN for encoding body-part features into a human-based spatio-temporal graph. This proposed methodology includes a light-weight adaptive GCN and exhibits excellent results on benchmark datasets.

Graph Sparsification and Redundancy Removal: Longer Spatio-temporal GCNs face a problem of trivial edges and redundant dependencies between different neighboring nodes. In order to overcome this problem, Liu *et al.* [54] devised a multiscale aggregation scheme using a spatio-temporal convolutional operator called as G3D. The G3D operator figures out the importance of nodes in different neighborhoods for effective long-range modeling. The proposed multiscale aggregation scheme is powerful for removing redundant dependencies between distant and adjacent neighborhoods. Ahmad *et al.* [47] worked on a similar idea of removing redundant and trivial edges in ST-GCN using edge-effective resistance. This concept of graph sparsification using edge-effective resistance is an extension of ST-GCN [22], which involves graph Laplacian and assumes that two graphs G and H are considered to be equal, if their Laplacian quadratic forms are identical, denoted as

$$k_1 * \Delta_H \leq \Delta_G \leq k_2 * \Delta_H \quad (7)$$

where k_1 and k_2 are constants. The authors evaluated their method on five benchmark datasets and proved this method state-of-the-art. Instead of sparsifying nodes and edges as mentioned above, another work by sparsifying skeleton frames in ST-GCN was proposed by Gao *et al.* [55]. This method can formulate a graph regression GCN for skeleton-based action recognition, where graph regression can statistically learn the underlying graph observations. In this technique, the authors implemented spectral GCN and proposed a sparsified spatio-temporal graph for each frame, its previous frame and for subsequent frame via graph regression. The authors evaluated their method on three benchmark datasets.

Graph Reasoning: In the graph theory literature, graph-based reasoning is an important area of research for modeling relational structures. Ding *et al.* [56] were the first to exploit

relational reasoning for the body-parts of skeleton sequence. In this method, the skeleton sequence is treated as a spatio-temporal graph; subsequently a network is devised for modeling graph interactions via spatio-temporal graph interaction networks. The model takes a spatio-temporal graph as input and relational reasoning is performed as rigid bodies. The main contribution of this article is to establish a relationships between the disjoints and distant joints of the body parts. This technique can provide long-term temporal information for a skeletal sequence and the efficacy of model is higher for three large-scale benchmark datasets. Another work using graph-based reasoning was carried out by Si *et al.* [57] by modeling a spatio-temporal evolution using hierarchical spatial reasoning and temporal stack learning. The hierarchical spatial reasoning network is responsible for capturing the intraspacial and body structural information for each part. On the other hand, the temporal stack learning network captures the temporal dynamics of body skeleton and includes multiple skip-clip LSTM. The authors proposed a novel clip-based incremental loss to improve the performance of stack learning. The clip-based incremental loss is defined as

$$L = - \sum_{m=1}^M (m/M) \sum_{i=1}^C (y_i) \log(\hat{y}_{mi}) \quad (8)$$

where y_i is the actual output, \hat{y}_{mi} is the predicted output, and C is the number of classes. M denotes the total number of video clips. Graph-based reasoning is a powerful concept, therefore, Zhang *et al.* [58] attempted to model a long-term temporal relationships between video sequences. In this article, it is proposed a temporal reasoning graph (TRG) to capture the appearance features and temporal relations. A multihead temporal relation aggregator was proposed for extracting the semantic relationship between graphs. The main contribution of this article was the development of a TRG and multihead graph representation. The experiments were performed on activity datasets and where this TRG method performed state-of-the-art.

VideoGraphs: In abovementioned ST-GCN methods, body skeleton is taken as the input to GCN, however, Wang *et*

al. [59] introduced a novel video representation as a space-time region graph. In this representation, graph nodes symbolize object region proposals belonging to different frames. Graph nodes reflect the similarity between correlated objects and the interaction between corresponding objects is determined using spatio-temporal graphs. This modeling of human-object and object-object interactions using spatio-temporal graphs was the key idea of this article. The authors evaluated their model on challenging activity recognition datasets, where it showed the best performance.

The earlier ST-GCN architectures were facing some drawbacks, which are subsequently overcome in the later designs. One of the main disadvantages of pioneering ST-GCN is that it employs the fixed topology of graph as spatio-temporal graph, which only represents the physical structure of human body. For example, “clapping” or “dancing” actions, the two hands move together and have some relationship, but ST-GCN does not provide any relationship between these two hands while performing these actions. Second, ST-GCN only includes first-order information (spatial coordinates of joints and bones), but does not include any second-order information (bones lengths or directions, etc.)

B. Graph Recurrent-Attention Network

Recurrent and recursive architectures take the output of previous time step as the input of current step for modeling sequential data [60]. This categorization includes attention-mechanism with GCN for emphasizing the important parts of the body skeleton, pertaining to some action. Such recurrent-attention methods include pooling mechanism for GCN. A schematic illustration of attention-based GCN is shown in Fig. 3(b). The graph attention network bears two advantages that it could emphasize certain joints or bones using spatial attention and could also emphasize certain frames using temporal attention.

Attention GCN: An attention-based graph convolution network was proposed by Li *et al.* [61], with the intuition that all body joints are not equally important for action recognition. This fully end-to-end action-attending network can model irregular skeletons as undirected attribute graphs, using recurrent temporal encoding. The model is robust to adaptively detect various salient action units for different actions. The authors evaluated the performance of their model for medium and large-scale datasets. For skeleton-based action recognition, different joints make discriminant contributions to action recognition. To exploit this concept, Ahmad *et al.* [62] introduced spatial attention in order to investigate attention-joints for skeleton-based action recognition. In this method, only the features corresponding to attention-joints were computed and propagated for action recognition. There are the following three types of features corresponding to these attention-joints:

- 1) distances of attention-joints from the center of body;
- 2) distances between the neighboring attention-joints;
- 3) the joint-flow among consecutive frames.

From layer l to $l + 1$, message-passing relationship for this method is defined as

$$H(l+1) = \sigma(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)}). \quad (9)$$

Chen *et al.* [63] introduced the idea of structural graph pooling (SGP) with joint-wise channel attention. SGP incorporates prior knowledge of human body for the pooling of body skeleton graphs. This methodology is a good option for global representations and reduces the computational cost by decreasing computational parameters. SGP gradually expands graph convolutional receptive field using the pooling of human skeletons. It is implemented in a gradual manner, where the first SGP layer has 25 nodes, which are gradually pooled out to ten nodes, and then pooled out to 5-nodes (left-hand nodes, right-hand nodes, and trunk nodes). To efficiently extract discriminative spatial and temporal features for action recognition, an attention enhanced graph convolutional LSTM network (AGC-LSTM) was proposed by [64]. This technique enhances the temporal receptive field using temporal hierarchical architecture. An attention mechanism is embedded in each AGC-LSTM block to identify the key joints. Additionally, each AGC-LSTM block contains temporal average pooling and feature augmentation (FA). For AGC-LSTM, attention-score for the nodes is calculated as follows:

$$\alpha_t = \sigma(U_s \tanh(W_h \hat{H}_t + W_q q_t + b_s) + b_u) \quad (10)$$

where $\alpha_t = (\alpha_{t1}, \alpha_{t2}, \dots, \alpha_{tN})$, and U_s , W_h , W_q represent the learnable parameter matrices. b_s and b_u represent the bias-parameters. q_t is a query feature. This concept of AGC-LSTM is also illustrated in Fig. 6. For simultaneous modeling the dependency of physically connected and disconnected joints, a new motif-based GCN was proposed for skeleton-based action recognition by Wen *et al.* [65]. This definition of a motif (or metagraph) has been borrowed from [66] and [67] as a pattern of connections among different node types. Motif-based GCN includes a variable temporal dense block for local temporal information; meanwhile, nonlocal blocks use attention mechanism for capturing global temporal dependencies. Tian *et al.* [68] investigated the attention mechanism and co-occurrence feature learning for skeleton-based action recognition using graph convolution. The proposed architecture employs the attention branch to emphasize the most significant features, and co-occurrence features are globally aggregated. The inherent correlation is exploited in various branches to improve the classification accuracy and convergence speed. For this method, attention mechanism is implemented using fully attention blocks. Huang *et al.* [69] devised a part-level GCN that can capture the information from body skeleton at different part levels. The part-level method is implemented using part-relation block and part attention block. The part relation blocks bring out the high-level relationships between different body-parts, whereas part attention blocks assign different level of attention to different body parts. These two blocks work in a complementary manner (called as part-level GCNs), and exhibits the state-of-the-art performance. Yang *et al.* [70] raised a very important issue that adjacency matrix of the body skeleton is fixed and only models the intrinsic relationship between joints and bones. While performing different actions, distant nodes may function together and generate an extrinsic relationship between distant joints, e.g., the clapping-hands action involving the movement of two hands, with distant joints. To establish this distant relationship

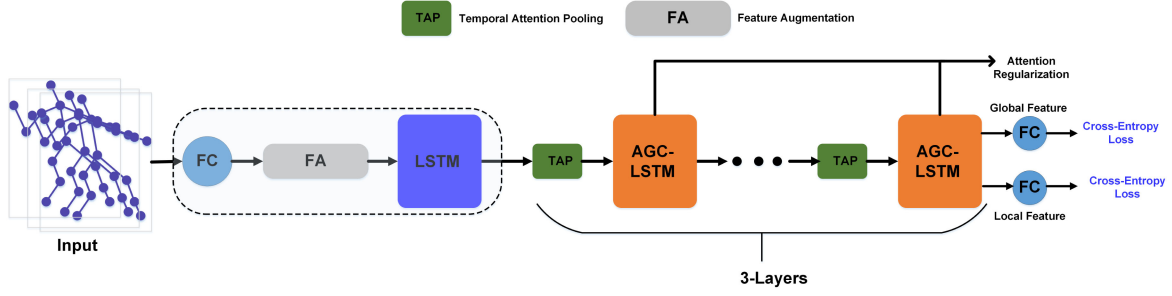


Fig. 6. Illustration of AGC-LSTM [64]. FA computes feature differences with position features and concatenates both position features and feature differences. LSTM is used to dispel scale variance between feature differences and position features. Three AGC-LSTM layers can model discriminative spatial-temporal features. Temporal average pooling is the implementation of average pooling in the temporal domain.

between joints, a pseudo-GCN with temporal and channel-wise attention was proposed [70]. A learnable matrix replaces the fixed adjacency matrix. The authors evaluated the performance of this model on benchmark datasets and proved it state-of-the-art. To draw the dependencies between joints, Shi *et al.* [71] propose the decoupled spatio-temporal attention network for action recognition. Such attention-blocks are good for modeling spatial-temporal dependencies between joints. These techniques proposed for attention-block encompass spatial-temporal attention decoupling, decoupled position encoding, and spatial global regularization. The authors tested their model on four benchmark datasets for the performance comparison. In order to alleviate the overfitting and computational burden in GCN-based action recognition, Cheng *et al.* [72] introduced attention-guided DropGraph technique. DropGraph is proposed as a regularization method for graph convolution in order to discard nodes redundant features. Moreover, authors also come up with a decoupling GCN in order to boost-up graph modeling for industrial development. The authors evaluated their model on benchmark datasets and established the superiority of this method. Feng *et al.* [73] studied new spatial and view attention modules for view-invariance and noise suppression in skeleton-based action recognition. The authors modeled a view-specific subnetwork for learning discriminative features from different input views. This technique generates a view-invariant representation of input skeleton, which is fed to graph network for recognizing different actions.

LSTM-Recurrent GCN: Recurrent graph networks are excellent choice for modeling temporal information for action recognition. Zhang *et al.* [74] introduced graph convolutional operations with LSTM architectures for modeling spatio-temporal features for action recognition. The graph convolution operation is introduced in a recurrent manner at each time step for input-to-state and state-to-state transitions. The proposed GC-LSTM framework is effective for determining discriminative frames and is insusceptible to irrelevant details for action recognition. This method implemented convolution operation in spectral domain; authors conducted their experiments on medium and large-scale datasets and the proposed method exhibited the state-of-the-art performance. Huang *et al.* [75] proposed a new concept of long-short graph memory network (LSGM) for extracting latent structural dependency between nodes. This method employs a

module called graph temporal spatial calibration (GTSC) for calibrating temporal-spatial features. LSGM combined with GTSC (LSGM + GTSC) resulted in the best performance, proving to be state-of-the-art module. Bidirectional LSGM with a temporal attention module, was used mainly for key-frame selection. Zhao *et al.* [76] investigated structure-aware feature representation using the LSTM network for modeling the temporal dynamics of data. This model embraces Bayesian probabilistic modeling for capturing statistical features in the data. An adversarial prior is used for regularizing and generalizing the model. The empirical results on benchmark datasets indicated that proposed method produces the best performance. The message-passing for Bayesian GC-LSTM is defined as

$$H^{(l+1)} = \sigma \left(\sum_{k=1}^K (M_k \circ F_k) H^{(l)} W_k^{(l)} + b_k^{(l)} \right) \quad (11)$$

where F_k is a symmetric matrix known as, $F_k = \tilde{D}^{-1/2} \tilde{A}_k \tilde{D}^{-1/2}$, M_k is a learnable mask computed using Bayesian inference. Considering the fact that different spatial parts of the body skeleton can play distinctive roles in different temporal phases of action recognition. Yang *et al.* [77] introduced an impressive method for enhancing spatial and temporal receptive fields using multineighborhood graph convolution with multiple LSTM networks having different time windows. The graph representation ability is enhanced via an attention-gate with LSTM modules and various levels of temporal dynamics are extracted using multiscale LSTMs. This method is idiosyncratic in a sense that convolution operation for this method is implemented in spectral-domain.

Video Interaction Graphs: Recognizing human activities in long-temporal videos has always been a challenge; therefore, Hussein *et al.* [78] proposed a model to recognize minute-long videos using video graphs. It was established that long videos are comprised of small video actions and their interacting objects, which can be inter-related using video graphs. VideoGraph consists of small graphs where a strong relationship between different videos or interacting objects is established as strong edges. Graph nodes in the network behave deterministically whereas graph edges behave in a probabilistic manner. A node attention architecture that can highlight the most promising nodes for action recognition is devised. This architecture can recognize

actions in 30-min long videos; however, the best performance was achieved for 1-min long videos. Group activities, such as sports have discriminant levels of interactions between different subjects; therefore, in order to model this discriminant interaction, Lu *et al.* [79] proposed a graph attention and interaction model (GAIM). GAIM includes graph attention block which explicitly models the unbalanced interaction between different subjects of an activity. The spatial structures and semantic appearances of the scene are modeled as spatio-temporal graphs, which are further excited to GCNs having a shared attention mechanism. The authors conducted their experiments on four publicly available activity datasets and reported the state-of-the-art performance. The authors claimed that this was the pioneering study using GCN for group activity recognition. Qi *et al.* [80] proposed an attention-based semantic RNN for analyzing group activities. The authors introduced two types of attention: spatial and temporal attention. This technique involves a semantic graph, which is modeled for describing spatial context using a structural-RNN. This proposed spatio-temporal attention model determines the key persons and frames for the improved performance.

In the GCN literature, it is mentioned a potential shortcoming of the attention-recurrent network that receptive fields of these methods are inflexible and may result in lesser expressiveness [81].

C. Two-Multistream GCN

Multistream GCN includes two or more streams of graph convolution networks for action recognition. Each stream is excited with different inputs (e.g., joints or bones) and then features in each stream are emphasized. A schematic diagram for this method is shown in Fig 3(c). Two and multistream techniques are topology-learnable and data-driven methods, which bring flexible graph structure. The other main advantage of this categorization is that it could involve second-order information, e.g., length of bones and directions.

Two-Stream GCN: Using first- and second-order information of joints and bones, Shi *et al.* [82] proposed a novel two-stream adaptive GCN (2s-AGCN) for skeleton-based action recognition. The authors used bone-stream and joint-stream as two-stream framework, which is effective for simultaneously modeling the first- and second-order information and promises a considerable increase in action recognition accuracy. In this method, the topology of graph is optimized together with other parameters of the network in an end-to-end learning manner. The 2s-AGCN is mathematically defined as

$$H^{l+1} = \sum_k^{K_v} W_k^l H^l (A_k + B_k + C_k). \quad (12)$$

Here, the adjacency matrix includes the following three matrices: A_k , B_k , and C_k . A_k is the original $N \times N$ adjacency matrix. B_k is also $N \times N$ adjacency matrix, learned during training process. C_k is $N \times N$ node similarity matrix. The two-stream architecture is the pioneering work, which provided a strong foundation for GCN-based action recognition using the two-multistream method [83]–[86], etc. This two-stream

architecture is shown in Fig. 7. In contrast to fixed graph structures of ST-GCN in [22], 2s-AGC entails data-driven optimal graph structures for different action classes. Another two-stream network was proposed as node and edge graphs by Li *et al.* [83]. This method incorporates the following three kinds of partition strategies:

- 1) unlabeled;
- 2) distance partitioning;
- 3) spatial configuration partitioning.

In contrast to two-stream architecture proposed in [82], graph edge representation is made by formulating the centers of bones. The node-stream and edge-stream are fused together, then pooling and softmax functions are used to obtain the prediction probability for each class. Luo *et al.* [84] proposed a novel reasoning interaction-based GCN. This two-stream framework can model the interaction between objects and scene patches. A CNN stream extracts object features, while discriminative scene detector models the scene patches. Then, interaction between objects and scenes is modeled as a graph, which is fed to GCN for class prediction. Shi *et al.* [85] represented skeleton data as a directed acyclic graph, determining the kinematic dependency between joints, bones, and their relationship. The topological structure of a graph is adaptive during training, which significantly enhances the performance. This technique operates in two-stream manner, which includes the motion of joints and deformations of bones for action recognition. The output of two-stream network is fused using a softmax-layer for final classification.

A seminal work using zero-shot learning with two-stream GCN was carried out by Gao *et al.* [86]. In this article, it is formulated knowledge graphs for modeling explicit relationships using zero-shot action recognition. Such structured knowledge graphs can model relationships between action–attributes, action–action, and attribute–attribute. The proposed method works as a two-stream GCN, having classification branch and instance branch. The authors validated the results of this method on benchmark datasets. Classifier branch takes subject and scene information as semantic-embedding vectors for action recognition, whilst instance branch models attribute embedding. This proposed method works in an end-to-end fashion such that generated classifiers and instance features coadapted and cooperated for final classification. Gao *et al.* [87] extended the idea of [86] by using word embedding and zero-shot learning. In this article, a task-driven message passing method is introduced, called prototype-sample GCN, having a prototype branch and a sample branch. Prototype branch deals with learning the representation of video categories, while sample branch generate video samples by leveraging object semantics. The authors evaluated their method on five benchmark datasets and established that proposed framework can perform better than contemporary methods. Exploiting the idea of knowledge transfer between different graphs, Tang *et al.* [88] proposed the idea of graph interaction networks (GIN), which exploits knowledge transfer between two graphs. GIN emphasizes higher level representation using a self-learned weight matrix. This method functions in a two-stream fashion as source-domain and target domain; where the GIN shares weights between two

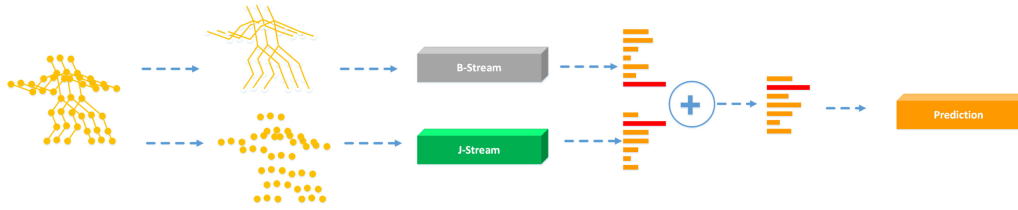


Fig. 7. Illustration of overall architecture of 2s-AGCN [82]. The scores of two streams are fused to obtain the final prediction.

domains. In order to encompass the information of scene objects, Shi *et al.* [89] proposed a novel two-stream graph convolutional architecture with scene object information for skeleton-based action recognition. The scene objects contain abundant pragmatic information for action recognition, where objects are modeled as a nodes. A relationship is established between each object node and intrinsic body skeleton. The model functions as a graph skeleton having a joint-stream and a limbs-stream. The authors evaluated their model on action recognition benchmark dataset.

Multistream and Multiscale GCN: A work using higher order spatial and temporal features for skeleton-based action recognition was carried out by [90]. Higher order features, such as the velocities of joints and bones, Euclidean distances among 3-D joints and accelerations of joints and bones were included. These higher order features are aggregated in a multistream fashion and the softmax function is applied for final classification. To recognize body actions in noisy and incomplete skeletons, Song *et al.* [91] formulated a multistream GCN for such skeletons. This GCN includes class activation maps and is denoted as richly activated GCN, where richly discovered features improve the robustness of model. Another multistream (involving six modalities) graph convolutional architecture was introduced by Li *et al.* [92]. This framework involves six modalities: joints, bones, their motions, and relative positions. The features of these six modalities are serially connected as spatio-temporal graph convolutional layers. The authors evaluated their proposed method on two large-scale action recognition datasets. In order to enrich multilevel information, Li *et al.* [93] developed a multiscale and multistream improved-GCN. This approach involves an improved-GCN featuring blocks with multiple parallel kernels. This GCN architecture includes a deep architecture of ten layers and residual connections between layers. This method uses five multistreams: joints, joint motions, bones, bone motions, and bones-cosine transform. The authors claim that bones-cosine stream contains important information and is very helpful for action recognition. Liu *et al.* [94] argues that GCNs overlook structural information for final classification because they employ fully connected layers and softmax layers. Therefore, authors propose a solution to this problem by involving a novel GCN-hidden conditional random field. This proposed model endeavors to retain the structural information of body skeleton for final classification. The authors designed a multistream network with following three streams: the joint-coordinates, bone direction, and temporal displacement of joints between consecutive frames. The authors tested their

model on benchmark datasets and it exhibited the state-of-the-art performance. A centrality GCN for skeleton-based action recognition was recently investigated by [40]. The authors attempted to bring forth overlooked topological information for skeleton-based action recognition using the centrality of joints, bones, and subgraphs in a multistream manner. Joints centrality indicate the relevance between nodes in the spatial domain; bone centrality determines the degree of importance of edge in a graph, and the subgraph centrality builds a connection between disconnected body parts for action recognition. This proposed centrality GCN is based on spectral convolution using Chebyshev polynomials. Motion prediction and 3-D skeleton-based action recognition have been studied separately for action recognition, however, their internal correlations for action recognition have been overlooked. Therefore, Li *et al.* [95] devised a symbiotic-GCN for capturing the motion of body-joints and parts for action recognition. The backbone of this method is a multibranch, multiscale GCN for extracting spatial and temporal features, where a multiscale GCN is applied as a joint-scale and part-scale graph. The joint-scale graphs handle actional graphs for action-based relations and structural graphs to determine the physical constraints. In contrast, part-scale graph convolution (PGC) exhibits an idiosyncrasy for extracting multiscale spatial information. The main contributions of this article are joint-scale graph convolution and PGC operators for extracting multiscale spatial information. The temporal convolution is modeled using multiordered motion differences, reflecting the positions, velocities, and accelerations of joints.

D. Encoder-Decoder GCN

This categorization yields an inherent low-rank representation of body skeleton using graph-encoder. Then decoder networks produce low-ranked graphs with strong features for GCN-based action recognition. Li *et al.* [96] proposed an encoder-decoder framework for capturing stronger dependencies for skeleton-based action recognition. Actional-links are adept for modeling action-specific latent dependencies. Higher-order dependencies are modeled by using structural links. This proposed method is termed as actional-structural inference model because it exploits the actional (extrinsic) relationship between joints and structural (intrinsic) relationship between the joints. The authors validated this methodology on two skeleton datasets and the proposed method performed significantly well over other methods. Motion prediction of 3-D body skeleton can also be helpful for recognizing human actions; therefore, Li *et al.* [97] proposed a dynamic

multiscale GCN. This model works in an encoder–decoder fashion; the encoder comprises of multiscale graph convolutional units, and the decoder generates future poses by using graph-based gate recurrent units. Although 3-D convolutional networks are effective for modeling spatio-temporal information, but they suffer from the long temporal kernels. Action segmentation and localization can be very helpful for recognizing different actions, therefore, Ghosh *et al.* [98] studied a stack hourglass architecture for action segmentation. This proposed architecture works in an encoder–decoder design and thrives various kind of descriptors such as frame-level VGG, segment-level I3D for joints inference. The empirical experiments are carried out on activity datasets.

During this research, it is noticed that nominal work has been done using encoder–decoder GCN for action recognition, this area has yet to be explored further for underlying capabilities of encoder–decoder architecture for action recognition.

E. Miscellaneous GCN Techniques

Miscellaneous methods for GCN do not fall into the aforementioned categories of GCN-based action recognition, and include a wide variety of graph neural architectures, e.g., reinforcement learning, generative-adversarial design, or few-shot learning, etc. In this categorization, the proposed methods are quite naive and in developing stage for GCN-based action recognition. Tang *et al.* [99] emphasized that different frames in a video contain discriminant levels of information for action recognition; therefore, if it can be identified such high-level information frames, it can be accurately recognized human actions using only the significant frames. In this regard, authors established a pioneering work by proposing a deep progressive reinforcement learning method for action recognition that can identify the most informative frames and discard ambiguous and redundant frames in a video sequence. Deep reinforcement learning is used in a progressive manner for modeling frame selection by considering: 1) the quality of the selected frames and 2) the relationship between selected frame and the whole body. In this architecture, a GCN network is used in final stage to capture joint dependencies for action recognition. In order to alleviate the problem of fixed graph architecture, Li *et al.* [100] formulated a dynamic structure-induced intra- and inter-GCN for action recognition. The intragraph models features of different body parts as a graph, and then intergraph models the relationships between these different body parts (intragraphs). The proposed framework is end-to-end trainable and the intrapart and interpart graph accurately capture the internal modalities of different body parts. Zhang *et al.* [101] devised a graph edge convolutional network corresponding to bones in the body skeleton. The edges in body skeleton are redefined as their neighboring spatial edges. Since the nodes and edges are complementary, this proposed method investigates graph node convolution along with the graph edge convolution. The proposed architecture exhibited promising results and the state-of-the-art performance for benchmark datasets. It is established that graph convolution is well-defined in the spectral-domain

for capturing the topology of manifolds and data. The key factor for the success in spectral-domain GCN is owing to Laplacian operators. Therefore, exploiting this concept, Mazari *et al.* [102] proposed a multi-Laplacian GCN that can learn graph-Laplacian as convex combinations of elementary Laplacians. Additionally, the authors introduced a novel pooling operator for graphs that includes context-dependent node expansion with global average pooling.

The GCN network is considered to be computationally expensive, requiring ≥ 15 G-FLOPs (Giga Floating-number Operations). To overcome the computational complexity of GCN-based methods, Cheng *et al.* [81] proposed a novel idea of shift GCN, which involves lightweight point-wise convolutions. Empirically, it is demonstrated that the proposed shift-GCN method can have ten-times less computational complexity and comparable performance for three action recognition benchmark datasets. Despite the spatial distances between joints, some distant joints are important for skeleton-based action recognition. Therefore, a context-aware GCN was proposed by Zhang *et al.* [103], which incorporates the contextual information of corresponding joints for skeleton-based action recognition. Context-aware network is applied in first stage in order to figure out the relationship and interdependencies between joints, then GCN is applied for the classification of such context-aware graphs. Zhu *et al.* [104] proposed a topology-learnable GCN for skeleton-based action recognition. The authors attempted to model the self-learning ability of GCNs, through feature learning and graph vertex fusion. Topology-learnable GCN is based on the architectures of ST-GCN [22] and 2s-AGCN [82], with the same number of layers and channels. The proposed methodology is validated on two large-scale action recognition benchmark datasets.

In [105], a novel technique for matching skeleton graphs was developed by incorporating a pairwise graph similarity measure. This method represents videos as skeleton graphs of the foreground subjects. Such representation is considered to be insensitive to illumination variations, subject appearance, and background. The proposed method is evaluated on small and medium-scale action recognition datasets. Guo *et al.* [106] proposed a novel method for recognizing unseen 3-D actions using only a few examples, termed as neural graph matching (NGM). This method uses a graphical representation for action recognition by exploiting the inherent structure of 3-D data. NGM functions in an adversarial manner where the graph generator and graph matching work in an end-to-end fashion. The node matching is the first step toward graph matching, and therefore, node matching feature $f(\hat{\cdot})$ for a node x_i is defined as

$$f(\hat{x}_i) = \sigma(\tilde{D}^{-1/2} A_i \tilde{D}^{-1/2} f(x_i) W_{\text{edge}}) \quad (13)$$

where W_{edge} is the edge-weight matrix. For two interaction graphs G_i and G_j , the graph matching metric is defined as

$$\phi_{GM}(G_i, G_j) = \|T(G_i) - T(G_j)\|^2 \quad (14)$$

where $\phi_{GM}(G_i, G_j)$ is the distance between corresponding graph tensors $T(G_i)$ and $T(G_j)$.

TABLE II
SUMMARY OF COMMONLY USED DATASETS

Dataset	Year	# Classes	# Subjects	# Videos	# Joints	Papers
NTU-RGBD	2016	60	40	56,880	25	[95], [82], [91], [85], [22] [45], [77], [54], [90], [69] [64], [68], [63], [65], [75] [62], [103], [81], [99], [101] [46], [96], [76], [51], [55] [48], [52], [50], [92], [61] [88], [83], [93], [94], [100] [74], [57], [56], [104], [70] [40], [47], [53], [71]
Kinetics-Skeleton	2017	400	-	266,440	18	[95], [82], [85], [22], [45] [77], [54], [68], [63], [65] [103], [101], [96], [51], [52] [89], [92], [56], [104], [40] [47]
SYSU-3D	2015	12	40	480	20	[69], [63], [75], [99], [76] [55], [94], [57]
NTU-RGBD-120	2019	120	53	114,480	25	[54], [90], [81], [93], [71]
Northwestern-UCLA	2014	10	10	1494	-	[64], [81], [94], [57]
UT-Kinect	2013	10	10	2000	-	[99], [55], [74]

A common disadvantage for abovementioned GCN-based action recognition methods is that there is a lack of explicit visual information or object related information to solve such problems. Second major shortcoming is that computational complexity of most of GCN-based methods is very large, for example, the computational complexity for ST-GCN can reach 16.2 GFLOPs (Giga-Floating OPERations) for one action sample.

IV. DATASETS AND RELATED CODES

The action recognition task entails how to classify a human action with an accurate label for an image. Medium-scale (SYSU-3D, UT-Kinetics, and MSR Action) to large-scale and very challenging nanyang technology university - red green blue depth (NTU-RGBD) and Kinetics-Skeleton) skeletal-action datasets have been devised over a decade of development in the area of human action recognition. Table II presents a summary of these commonly used datasets for GCN-based action recognition.

A. NTU-RGBD

The NTU RGB+D dataset [107] is a multimodal dataset, with RGB, skeleton, and depth maps of body joints. The dataset contains 56 880 indoor action videos of 40 subjects with age groups ranging from 10 to 35 years, where 25 joints were extracted for each subject in X , Y , and Z coordinates. There are 60 different action classes, with cross-subject (CS) and cross-view (CV) evaluations in this dataset. The videos were simultaneously captured by three Microsoft Kinect v.2 cameras at angles of -45° , 0° , and 45° . For the cross-subject (X-Sub) setting the actor is different in training and testing; 40 320 samples are used for training and 16 560 samples are used for testing purpose. In the cross-view (X-View) setting, 37 920 samples from camera 2, and 3 are used for training, and 18 960 samples from camera 1 are used during the testing. Generally, authors report Top-1 accuracy for the X-Sub and X-view settings. Interestingly, as

indicated by Table II for most of GCN-based action recognition studies, the models are evaluated using the NTU-RGBD dataset.

B. Kinetics-Skeleton

Deepmind Kinetics human action dataset [108] contains approximately 300 000 YouTube action videos. The dataset contains 400-action classes with 400 videos for each action class. The length of each video clip is approximately 10 s. The original Kinetics dataset does not have any skeleton information. To obtain the 2-D coordinates of skeleton joints, Yan *et al.* [22] used OpenPose toolbox [109]. The OpenPose toolbox can predict 18 joints for each person, providing 2-D coordinates (X , Y) for the predicted joints in the image coordinate system and corresponding confidence score “C.” Therefore each joint is represented as (X , Y , C), and for more than one person in a frame, the person with lower confidence is ignored. The Kinetics skeletal dataset contains 266 440 video samples, where 246 534 samples are used for training and the remaining 19 906 samples are used for testing. It is general practice that authors report Top-1 and Top-5 accuracy for this dataset.

C. NTU-RGBD-120

Liu *et al.* [110] proposed a large-scale dataset having 114 480 videos of 120 different actions. This dataset is an extension of the NTU RGB+D-60 dataset, which contains a wide range of action classes including daily, mutual, and health-related activities. The 53 subjects who participated in this dataset vary between the age of 10 to 57#years. NTU-RGB+D-120 also includes the following four modalities: RGB, depth sequence, skeleton data, and infrared frames collected using Microsoft Kinect v2. The skeleton-data modality marks 25 joints on the human body. This dataset contains videos of 106 different subjects and, two different evaluation settings are used as cross-subject and cross-setup evaluations.

D. SYSU-3D

Hu *et al.* [111] devised a human activity SYSU-3D dataset, having 480 videos of 40 different subjects engaging in 12 different activities. The dataset contains 30 coordinates from 20 different joints, associated for each frame of the sequence. There are two settings available for this dataset, for setting 1, half of the video samples are used for training while remaining half are used for testing purposes. For setting 2, half of the actors' videos are used for training whereas remaining half of the actors' videos are used for testing.

E. Northwestern-UCLA

Wang *et al.* [112] created a video dataset having 1494 videos of ten different actions. This dataset was prepared by simultaneously capturing the videos from three Kinect cameras. Each video sequence contains both RGB and skeleton data. The videos from two cameras are used for training set, and the videos from third camera are used for testing set.

F. UT-Kinect

Hussein *et al.* [113] devised a skeleton dataset with 200 video sequences having 20 body joints per frame. The dataset includes ten classes of action datasets, which were performed by ten different subjects. Mainly, the authors used leave-one out cross-validation method for evaluation.

G. MSR Action3D

Li *et al.* [114] presented two-person interaction multimodal action recognition datasets, which had 20 action classes. There are total 557 video actions performed by ten different subjects.

H. Papers With Code

In Table III, we list the papers, which are provided with their source codes for reimplementation or a better understanding of different methods. We also compared the performance of contemporary action recognition methods for NTU-RGBD and Kinetics datasets in Table III. Most of the GCN architectures were implemented in Pytorch-Geometric, which is developed by Fey *et al.* [115]. Pytorch-Geometric is the Pytorch implementation for the graph neural network and is also accompanied by web resources, guidelines, and tutorials. Deep graph library is the other famous library for the graph neural network implementation [116]. Some recently developed graph neural network libraries are TensorFlow Graph Nets library [117], [118] and jraph [119], both developed by Deepmind.

V. DISCUSSION AND FUTURE DIRECTIONS

Over a short period, GCN has achieved considerable success for action recognition. In the literature, convolutional neural networks were incorporated for action recognition because they excellently model spatial information [2]. Additionally, temporal information has been embedded either through optical flow [8] or by 3-D convolutional neural networks [120], [121] for action recognition. Another methodology for action recognition

includes RNN or LSTM architectures [122]. While considering that GCNs have intrinsic capability for concurrently modeling the spatio-temporal body skeleton over the length of the video sequence [22]. This co-occurrence of spatio-temporal skeleton information makes the GCN a superior method over CNN or RNN (or a combination of both). It is observed that recent research trend has turned toward GCN and Table IV indicates that studies are increasingly being performed for action recognition using GCN. A performance comparison for the CNN, RNN, and GCN methods is also presented in Table IV.

We anticipate that recent advancements in GCN-based action recognition are built on strong conceptualization and development in GCN methods and domain knowledge from action recognition. For instance, it was firstly devised the idea of GCN by [21], which was incorporated in action recognition using ST-GCN [22]. Similarly, the graph attention network [123] was introduced and applied for action recognition, discussed in Section III-B. In line with this intuition, a two-stream network [2] was originally formulated to explore the domain knowledge of action recognition, which was later investigated for GCN-based action recognition methods, as explained in Section III-C.

Since the research trend is going to significantly shift from CNN and RNN methods toward GCN methods, therefore, it can be anticipated that a lot of potential and future works are underway to further investigate GCN for better action recognition.

RGB Fusion with GCN: It is imperative to combine the RGB modality with skeleton data for GCN-based action recognition [85], [90]. The RGB modality contains important information and its fusion with the skeleton modality must be investigated in future study.

Integrating Contextual Information: Integrating contextual information (e.g., image scene, objects, the interactions) with skeleton data for GCN can yield considerable success for action recognition [22], [64], [90]. Additionally, a method must be developed to fuse objects and human-object interaction for graph-based action recognition.

Long Temporal Modeling and Biasness: It is expected that a long-term ST-GCN can perform better by modeling long spatio-temporal information [69]. However, incorporating long-term spatio-temporal information introduces the problem of trivial nodes, resulting in an overfitting problem. Thus, it is immensely needed to explore graph sparsification techniques for modeling long spatio-temporal graphs for action recognition. Moreover, it will also be highly appreciated to model the biasness for solving the class imbalance problems [49].

Residual or Dense Connections: It is established that residual neural networks and dense-network have marked great success in solving deep learning problems. It is necessary to introduce these concepts of residual and dense connection modeling to GCN-based action recognition [75].

Novel Skeletal Representation: The body structural representation [64], [75], [96] and actional representation [96] are the recently used representations of body skeleton. The structural representation corresponds to intrinsic body skeleton, whereas, actional representation corresponds to inter-relation of joints for performing different actions. In addition to these two skeletal

论文及其代码的整理在表3中。

复现

用了Pytorch-Geometric库

TABLE III
PERFORMANCE COMPARISON FOR CONTEMPORARY GCN METHODS FOR NTU-RGBD AND KINETICS DATASETS

Approach	Category	Inputs	Pooling	Conv. Type	Source-code	NTU-RGB+D CS/CV	Kinetics Top-1/Top-5
Spatial-temporal GCN, [22]	ST-GCN	$A, X_{v \times d \times T}$	✓	Spatial	✓	81.5/88.3	30.7/52.8
Spatio-temporal GCN, [48]	ST-GCN	$A, X_{v \times d \times T}$	×	Spatial	×	74.8/86.3	-/-
Graph Representation, [50]	ST-GCN	$X^{Vid}, X_{v \times d \times T}$	×	Spatial	×	82.6/89.6	-/-
Relational reasoning GCN, [56]	ST-GCN	$A, X^{Vid}, X_{v \times d \times T}$	✓	Spectral	×	85.8/88.7	36.4/59.8
Part-based GCN, [46]	ST-GCN	$A, X_{v \times d \times T}$	×	Spatial	✓	87.5/93.2	-/-
Sparsified Graph, [55]	ST-GCN	$A, X_{v \times d \times T \times M}$	✓	Spectral	×	87.5/94.3	-/-
Reasoning+Stack learning, [57]	ST-GCN	X^{Vid}	×	Spatial	×	87.7/94.4	-/-
Neural Searching, [51]	ST-GCN	$U^{Fourier}, X_{v \times d \times T}$	×	Spatial	✓	89.4/95.7	37.1/60.1
Residual+dense connection, [52]	ST-GCN	$A, X_{v \times d \times T}$	✓	Spatial	×	89.6/95.7	37.3/60.4
Effective Resistance+GCN, [47]	ST-GCN	$A, X_{v \times d \times T}$	✓	Spatial	×	91.3/96.9	36.6/54.8
Disentangle-Unifying GCN, [54]	ST-GCN	$A, X_{v \times d \times T}$	✓	Spatial	✓	91.5/96.2	38.0/60.9
Temporal reasoning graph, [58]	ST-GCN	X^{Vid}	✓	Spatial	×	-/-	-/-
Space-time graph, [59]	ST-GCN	X^{Vid}	✓	Spatial	×	-/-	-/-
Action-attending GCN, [61]	ATT-GCN	$A, X_{v \times d \times T}$	✓	Spectral	×	72.7/82.8	-/-
Spatial+View-attention, [73]	ATT-GCN	$A, X_{v \times d \times T}$	×	Spatial	×	79.7/-	-/-
Bayesian-LSTM GCN, [76]	LSTM-GCN	$A, X_{v \times d \times T}$	×	Spatial	×	81.8/89.0	-/-
Graph-conv. LSTM, [74]	LSTM-GCN	$A, X_{v \times d \times T}$	×	Spectral	×	83.9/92.3	-/-
Motif-based GCN, [65]	ATT-GCN	$A, X_{v \times d \times T}$	✓	Spatial	✓	84.2/90.2	-/-
LSGM, [75]	LSTM-GCN	$A, X_{v \times d \times T}$	✓	Spatial	✓	84.7/91.7	-/-
M-GC LSTM, [77]	LSTM-GCN	$A, X_{v \times d \times T}$	✓	Spectral	×	85.3/93.8	32.3/55.2
Co-occurrence learning, [68]	ATT-GCN	$A, X_{v \times d \times T}$	✓	Spatial	×	86.8/92.1	32.8/55.2
SGP+JCA, [63]	ATT-GCN	$A, X_{v \times d \times T}$	✓	Spatial	×	86.9/93.8	-/-
PGCN-TCA, [70]	ATT-GCN	$A, X_{v \times d \times T}$	✓	Spatial	×	88.0/93.6	-/-
Part-level GCN, [69]	ATT-GCN	$A, X_{v \times d \times T}$	✓	Spatial	×	89.2/95.0	-/-
AGC-LSTM, [64]	ATT-GCN	$A, X_{v \times d \times T}$	✓	Spatial	×	89.2/95.0	-/-
Attention-Joint GCN, [62]	ATT-GCN	$A, X_{v \times d \times T}$	×	Spatial	×	90.7/-	-/-
DropGraph, [72]	ATT-GCN	$A, X_{v \times d \times T}$	✓	Spatial	✓	90.8/96.6	-/-
Decoupled-Attention GCN, [71]	ATT-GCN	$A, X_{v \times d \times T}$	✓	Spatial	✓	91.5/96.4	-/-
VideoGraph, [78]	ATT-GCN	X^{Vid}	×	-	×	-/-	-/-
Graph-Attention Interaction, [79]	ATT-GCN	X^{Vid}	✓	-	×	-/-	-/-
Stagnat, [80]	RNN-GCN	X^{Vid}, X^{Obj}	✓	-	×	-/-	-/-
Edge-node GCN, [83]	2s-GCN	$A, X_{v \times d \times T}$	✓	Spatial	✓	83.2/91.6	-/-
Richly-activated GCN, [91]	MS-GCN	$A, X_{v \times d \times T \times M}$	✓	Spatial	✓	85.9/93.5	-/-
2-stream GCN, [82]	2s-GCN	$A, X_{v \times d \times T}$	✓	Spatial	✓	88.5/95.1	36.1/58.7
Directed GCN, [85]	2s-GCN	$A, X_{v \times d \times T}, X_{e \times d \times T}$	✓	Spatial	×	89.9/96.1	36.9/59.6
Random Field+GCN, [94]	2s-GCN	$A, X_{v \times d \times T}$	✓	Spatial	×	90.0/95.5	-/-
Symbiotic GCN, [95]	2s-GCN	$A, X_{v \times d \times T}$	✓	Spatial	✓	90.1/96.4	37.2/58.1
Centrality GCN, [40]	MS-GCN	$A, X_{v \times d \times T}$	✓	Spectral	×	90.3/96.4	37.5/60.4
Multi-scale-stream GCN, [93]	MS-GCN	$A, X_{v \times d \times T}$	✓	Spatial	×	91.3/96.7	-/-
Multi-stream Enhanced GCN, [92]	MS-GCN	$A, X_{v \times d \times T \times M}$	✓	Spatial	×	91.4/96.8	39.4/62.1
High-order Feature+GCN, [90]	MS-GCN	$A, X_{v \times d \times T \times M}$	✓	Spatial	×	91.7/96.8	-/-
Object+GCN, [89]	2s-GCN	X^{Vid}	×	Spatial	×	-/-	32.1/55.0
Zero-Shot 2s-GCN, [86]	2s-GCN	X^{Vid}	×	Spatial	✓	-/-	-/-
Relation Model+Zero-shot, [87]	2s-GCN	X^{Vid}	✓	Spatial	✓	-/-	-/-
Graph Relation Interaction, [88]	2s-GCN	$X_{v \times d \times T}$	✓	-	×	-/-	-/-
Interaction Reasoning+GCN, [84]	2s-GCN	X^{Vid}	✓	Spatial	×	-/-	-/-

For NTU-RGBD dataset, we report performance for cross-subject (CS) and cross-view (CV) settings, and for Kinetics dataset, we list Top-1 and Top-5 accuracies.

TABLE IV
PERFORMANCE COMPARISON BETWEEN CNN, RNN, AND GCN TECHNIQUES
FOR NTU-RGBD DATASET

Method	Year	NTU-RGBD (%)	
		CS	CV
Visualization CNN, [35]	2017	76.0	82.6
Clips+CNN+MTLN, [4]	2017	79.6	84.83
Fine-To-Coarse 3D Action, [124]	2018	79.6	84.6
Trust-Gate ST-LSTM, [125]	2016	69.2	77.7
Two-stream RNN [126]	2017	71.3	79.5
Ensemble TS-LSTM, [127]	2017	74.6	81.3
VA-RNN, [128]	2017	79.4	87.6
Memory Attention Network, [129]	2018	82.6	93.2
Deep IndRNN, [130]	2019	86.7	93.4
Actional-Structural GCN, [96]	2019	86.8	94.2
2s-AGCN, [82]	2019	88.5	95.1
SGP-JCA, [63]	2020	86.9	93.8
Part-level GCN, [69]	2020	89.2	95.0
Neural Searching, [51]	2020	89.4	95.7
Disentangling-Unifying GCN, [54]	2020	91.5	96.2

representations, other novel representations are worth of investigation in future work.

Investigating Unsupervised Learning: Another possible future research direction is unsupervised or weakly supervised learning methods for GCN-based action recognition. We foresee that unsupervised or semisupervised methodologies could also lead to the way.

VI. CONCLUSION

In this article, we provide a comprehensive insight into different GCN techniques for action recognition. First, we devise a taxonomy to group different GCN techniques for action recognition into five subcategories. Then, we provide a detailed review of these methods, their mathematical models and contributions. A detailed description of various benchmark datasets along with corresponding source codes is presented. We also provide a

performance comparison for different techniques and conclude the survey by suggesting future research directions. To the best of the authors' knowledge, no such survey or insightful literature is currently available to describe the research developments in GCN-based action recognition. Therefore, we are hopeful that this article can serve the community and inspire the future research.

ACKNOWLEDGMENT

Mr. Tasweer Ahmad acknowledges to Chinese Scholarship Counsel (CSC) for the sponsorship of his Ph.D. degree.

REFERENCES

- [1] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles, "SST: Single-stream temporal action proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2911–2920.
- [2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Adv. Neural Informat. Process. Syst.*, vol. 27, 2014.
- [3] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.
- [4] Q. Ke, M. Bennamoun, S. An, F. Soheli, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3288–3297.
- [5] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with visual attention on skeleton images," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 3309–3314.
- [6] W. Zhu *et al.*, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016, pp. 3697–3703.
- [7] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," 2015, *arXiv:1507.02159*.
- [8] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.
- [9] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1390–1399.
- [10] J. Xu, K. Tasaka, and H. Yanagihara, "Beyond two-stream: Skeleton-based three-stream networks for action recognition in videos," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 1567–1573.
- [11] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, Aug. 2019.
- [12] S. Li, T. Jiang, T. Huang, and Y. Tian, "Global co-occurrence feature learning and active coordinate system conversion for skeleton-based action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 586–594.
- [13] M. Sheng, C. Liu, Q. Zhang, L. Lou, and Y. Zheng, "Vehicle detection and classification using convolutional neural networks," in *Proc. IEEE 7th Data Driven Control Learn. Syst. Conf.*, 2018, pp. 581–587.
- [14] A. El-Sawy, E.-B. Hazem, and M. Loey, "CNN for handwritten arabic digits recognition based on lenet-5," in *Proc. Int. Conf. Adv. Intell. Syst. Informat.*, 2016, pp. 566–575.
- [15] L. Tang and H. Liu, "Relational learning via latent social dimensions," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 817–826.
- [16] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1024–1034.
- [17] C. Zang, P. Cui, and C. Faloutsos, "Beyond sigmoids: The nettide model for social network growth, and its applications," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 2015–2024.
- [18] K. M. Borgwardt, C. S. Ong, S. Schönerauer, S. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, no. suppl_1, pp. i 47–i56, 2005.
- [19] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4438–4445.
- [20] F. Xu, J. Lian, Z. Han, Y. Li, Y. Xu, and X. Xie, "Relation-aware graph convolutional networks for agent-initiated social e-commerce recommendation," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 529–538.
- [21] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations*, 2017, pp. 1–14.
- [22] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.
- [23] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.
- [24] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," 2017, *arXiv:1709.05584*.
- [25] J. B. Lee, R. A. Rossi, S. Kim, N. K. Ahmed, and E. Koh, "Attention models in graphs: A survey," *ACM Trans. Knowl. Discovery Data*, vol. 13, no. 6, pp. 1–25, 2019.
- [26] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
- [27] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: A comprehensive review," *Comput. Social Netw.*, vol. 6, no. 1, pp. 1–23, 2019.
- [28] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Trans. Knowl. Data Eng.*, p. 1, 2020.
- [29] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent Kinect-based action recognition algorithms," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 15–28, Jul. 2019.
- [30] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," 2020, *arXiv:2002.05907*.
- [31] H.-B. Zhang *et al.*, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, 2019, Art. no. 1005.
- [32] H. A. Qazi, U. Jahangir, B. M. Yousuf, and A. Noor, "Human action recognition using sift and hog method," in *Proc. IEEE Int. Conf. Inf. Commun. Technol.*, 2017, pp. 6–10.
- [33] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.
- [34] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proc. ACM Multimedia Asia*, 2019, pp. 1–6.
- [35] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, 2017.
- [36] M. Meng, H. Drira, and J. Boonaert, "Distances evolution analysis for online and off-line human object interaction recognition," *Image Vis. Comput.*, vol. 70, pp. 32–45, 2018.
- [37] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: A benchmark for recognizing human-object interactions in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1017–1025.
- [38] D.-T. Le, J. Uijlings, and R. Bernardi, "TUHOI: Trento universal human object interaction dataset," in *Proc. 3rd Workshop Vis. Lang.*, 2014, pp. 17–24.
- [39] F. Patrona, A. Chatzitofis, D. Zarpalas, and P. Daras, "Motion analysis: Action detection, recognition and evaluation based on motion capture data," *Pattern Recognit.*, vol. 76, pp. 612–622, 2018.
- [40] D. Yang, M. M. Li, H. Fu, J. Fan, and H. Leung, "Centrality graph convolutional networks for skeleton-based action recognition," 2020, *arXiv:2003.03007*.
- [41] S. Tang, B. Li, and H. Yu, "ChebNet: Efficient and stable constructions of deep neural networks with rectified power units using chebyshev approximations," 2019, *arXiv:1911.05467*.
- [42] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, "CayleyNets: Graph convolutional neural networks with complex rational spectral filters," *IEEE Trans. Signal Process.*, vol. 67, no. 1, pp. 97–109, Jan. 2019.
- [43] M. Balciar, G. Renton, P. Héroux, B. Gauzere, S. Adam, and P. Honeine, "Bridging the gap between spectral and spatial domains in graph neural networks," 2020, *arXiv:2003.11702*.
- [44] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5115–5124.

- [45] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 8561–8568.
- [46] K. Thakkar and P. Narayanan, "Part-based graph convolutional network for action recognition," 2018, *arXiv:1809.04983*.
- [47] T. Ahmad, L. Jin, L. Lin, and G. Tang, "Skeleton-based action recognition using sparse spatio-temporal GCN with edge effective resistance," *Neurocomputing*, vol. 423, pp. 389–398, 2021.
- [48] L. Chaolong, C. Zhen, Z. Wenming, X. Chunyan, and Y. Jian, "Spatio-temporal graph convolution for skeleton based action recognition," 2018, *arXiv:1802.09834*.
- [49] B. Parsa, A. Narayanan, and B. Dariush, "Spatio-temporal pyramid graph convolutions for human action recognition and postural assessment," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1080–1090.
- [50] X. Gao *et al.*, "3d skeleton-based video action recognition by graph convolution network," in *Proc. IEEE Int. Conf. Smart Internet Things*, 2019, pp. 500–501.
- [51] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 2669–2676.
- [52] C. Wu, X.-J. Wu, and X. Kittler, "Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1740–1748.
- [53] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Action recognition via pose-based graph convolutional networks with intermediate dense supervision," 2019, *arXiv:1911.12509*.
- [54] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 143–152.
- [55] X. Gao, W. Hu, J. Tang, J. Liu, and Z. Guo, "Optimized skeleton-based action recognition via sparsified graph regression," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 601–610.
- [56] W. Ding, X. Li, G. Li, and Y. Wei, "Global relational reasoning with spatial temporal graph interaction networks for skeleton-based action recognition," *Signal Process.: Image Commun.*, vol. 83, 2020, Art. no. 115776.
- [57] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network," *Pattern Recognit.*, vol. 107, 2020, Art. no. 107511.
- [58] J. Zhang, F. Shen, X. Xu, and H. T. Shen, "Temporal reasoning graph for activity recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 5491–5506, 2020.
- [59] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 399–417.
- [60] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "Adding attentiveness to the neurons in recurrent neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 135–151.
- [61] C. Li, Z. Cui, W. Zheng, C. Xu, R. Ji, and J. Yang, "Action-attending graphic neural network," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3657–3670, Jul. 2018.
- [62] T. Ahmad, H. Mao, L. Lin, and G. Tang, "Action recognition using attention-joints graph convolutional neural networks," *IEEE Access*, vol. 8, pp. 305–313, 2019.
- [63] Y. Chen *et al.*, "Graph convolutional network with structure pooling and joint-wise channel attention for action recognition," *Pattern Recognit.*, vol. 103, 2020, Art. no. 107321.
- [64] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1227–1236.
- [65] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. Xia, "Graph CNNs with motif and variable temporal block for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 8989–8996.
- [66] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, 2016.
- [67] Y. Fang, W. Lin, V. W. Zheng, M. Wu, K. C.-C. Chang, and X.-L. Li, "Semantic proximity search on graphs with metagraph-based learning," in *Proc. IEEE 32nd Int. Conf. Data Eng.*, 2016, pp. 277–288.
- [68] D. Tian, Z.-M. Lu, X. Chen, and L.-H. Ma, "An attentional spatial temporal graph convolutional network with co-occurrence feature learning for action recognition," *Multimedia Tools Appl.*, vol. 79, pp. 12679–12697, 2020.
- [69] L. Huang *et al.*, "Part-level graph convolutional network for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11045–11052.
- [70] H. Yang, Y. Gu, J. Zhu, K. Hu, and X. Zhang, "Pgcn-tca: Pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition," *IEEE Access*, vol. 8, pp. 100 40–10047, 2020.
- [71] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial-temporal attention network for skeleton-based action recognition," in *Proc. Asian Conf. Comput. Vision*, 2020.
- [72] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling GCN with dropgraph module for skeleton-based action recognition," in *Proc. Eur. Conf. Comput. Vis.*, pp. 536–553, 2010.
- [73] Y. Feng, G. Li, and C. Yuan, "Spatio-temporal and view attention deep network for skeleton based view-invariant human action recognition," in *Proc. 11th EAI Int. Conf. Mobile Multimedia Commun. Eur. Alliance Innov.*, 2018, Art. no. 113.
- [74] H. Zhang, Y. Song, and Y. Zhang, "Graph convolutional LSTM model for skeleton-based action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 412–417.
- [75] J. Huang, Z. Huang, X. Xiang, X. Gong, and B. Zhang, "Long-short graph memory network for skeleton-based action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 645–652.
- [76] R. Zhao, K. Wang, H. Su, and Q. Ji, "Bayesian graph convolution LSTM for skeleton based action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6882–6892.
- [77] K. Yang, X. Ding, and W. Chen, "Multi-scale spatial temporal graph convolutional LSTM network for skeleton-based human action recognition," in *Proc. Int. Conf. Video, Signal Image Process.*, 2019, pp. 3–9.
- [78] N. Hussein, E. Gavves, and A. W. Smeulders, "Videograph: Recognizing minutes-long human activities in videos," 2019, *arXiv:1905.05143*.
- [79] L. Lu, Y. Lu, R. Yu, H. Di, L. Zhang, and S. Wang, "Gaim: Graph attention interaction model for collective activity recognition," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 524–539, Feb. 2020.
- [80] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. Van Gool, "Stagnet: An attentive semantic RNN for group activity recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 101–117.
- [81] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 183–192.
- [82] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12 026–12035.
- [83] G. Li, S. Yang, and J. Li, "Edge and node graph convolutional neural network for human action recognition," in *Proc. Chin. Control Decis. Conf.*, 2020, pp. 4630–4635.
- [84] W. Luo, C. Zhang, X. Zhang, and H. Wu, "Improving action recognition with the graph-neural-network-based interaction reasoning," in *Proc. IEEE Vis. Commun. Image Process.*, 2019, pp. 1–4.
- [85] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7912–7921.
- [86] J. Gao, T. Zhang, and C. Xu, "I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 8303–8311.
- [87] J. Gao, T. Zhang, and C. Xu, "Learning to model relationships for zero-shot video classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1, 2020.
- [88] Y. Tang, Y. Wei, X. Yu, J. Lu, and J. Zhou, "Graph interaction networks for relation transfer in human activity videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2872–2886, Sep. 2020.
- [89] X. Shi, H. Li, F. Liu, D. Zhang, J. Bi, and Z. Li, "Graph convolutional networks with objects for skeleton-based action recognition," in *Proc. IEEE Int. Conf. Ubiquitous Comput. Commun. Data Sci. Comput. Intell. Smart Comput., Netw. Serv.*, 2019, pp. 280–285.
- [90] J. Dong *et al.*, "Action recognition based on the fusion of graph convolutional networks with high order features," *Appl. Sci.*, vol. 10, no. 4, 2020, Art. no. 1482.
- [91] Y.-F. Song, Z. Zhang, and L. Wang, "Richly activated graph convolutional network for action recognition with incomplete skeletons," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1–5.
- [92] F. Li, A. Zhu, Y. Xu, R. Cui, and G. Hua, "Multi-stream and enhanced spatial-temporal graph convolution network for skeleton-based action recognition," *IEEE Access*, vol. 8, pp. 97757–97770, 2020.
- [93] W. Li, X. Liu, Z. Liu, F. Du, and Q. Zou, "Skeleton-based action recognition using multi-scale and multi-stream improved graph convolutional network," *IEEE Access*, vol. 8, pp. 144 529–144542, 2020.

- [94] K. Liu, L. Gao, N. M. Khan, L. Qi, and L. Guan, "A multi-stream graph convolutional networks-hidden conditional random field model for skeleton-based action recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 64–76, 2021.
- [95] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1, 2021.
- [96] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3595–3603.
- [97] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 214–223.
- [98] P. Ghosh, Y. Yao, L. Davis, and A. Divakaran, "Stacked spatio-temporal graph convolutional networks for action segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 576–585.
- [99] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5323–5332.
- [100] R. Liu, C. Xu, T. Zhang, W. Zhao, Z. Cui, and J. Yang, "Si-GCN: Structure-induced graph convolution network for skeleton-based action recognition," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.
- [101] X. Zhang, C. Xu, X. Tian, and D. Tao, "Graph edge convolutional neural networks for skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 3047–3060, Aug. 2020.
- [102] A. Mazari and H. Sahbi, "Human action recognition with multi-Laplacian graph convolutional networks," 2019, *arXiv:1910.06934*.
- [103] X. Zhang, C. Xu, and D. Tao, "Context aware graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14 333–14342.
- [104] G. Zhu, L. Zhang, H. Li, P. Shen, S. A. A. Shah, and M. Bennamoun, "Topology-learnable graph convolution for skeleton-based action recognition," *Pattern Recognit. Lett.*, vol. 135, pp. 286–292, 2020.
- [105] A. A. Abdulmunem, Y.-K. Lai, A. K. Hassan, and X. Sun, "Human action recognition using graph matching," in *Proc. AIP Conf. Proc.*, vol. 2144, no. 1, 2019, Art. no. 0 50003.
- [106] M. Guo, E. Chou, D.-A. Huang, S. Song, S. Yeung, and L. Fei-Fei, "Neural graph matching networks for fewshot 3d action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 653–669.
- [107] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB D: A large scale dataset for 3d human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1010–1019.
- [108] W. Kay *et al.*, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.
- [109] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
- [110] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan, and A. K. Chichung, "NTU RGB D 120: A large-scale benchmark for 3d human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.
- [111] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5344–5352.
- [112] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2649–2656.
- [113] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2466–2472.
- [114] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, 2010, pp. 9–14.
- [115] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch geometric," in *Proc. ICLR Workshop Representation Learn. Graphs Manifolds*, 2019, pp. 1–9.
- [116] M. Wang *et al.*, "Deep graph library: A graph-centric, highly-performant package for graph neural networks," 2019, *arXiv:1909.01315*.
- [117] P. W. Battaglia *et al.*, "Relational inductive biases, deep learning, and graph networks," 2018, *arXiv:1806.01261*.
- [118] Deepmind. *Graph Nets Library*. 2020. [Online]. Available: https://github.com/deepmind/graph_nets
- [119] Deepmind. *Jraph—A Library for Graph Neural Networks in Jax*. 2021. [Online]. Available: <https://github.com/deepmind/jraph>
- [120] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, "MiCT: Mixed 3d/2d convolutional tube for human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 449–458.
- [121] A. Diba *et al.*, "Temporal 3d convnets using temporal transition layer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 1117–1121.
- [122] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.
- [123] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.
- [124] T. M. Le, N. Inoue, and K. Shinoda, "A fine-to-coarse convolutional neural network for 3d human action recognition," 2018, *arXiv:1805.11790*.
- [125] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-Temporal LSTM With Trust Gates for 3d Human Action Recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 816–833.
- [126] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 499–508.
- [127] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1012–1020.
- [128] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2117–2126.
- [129] C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, and J. Chen, "Memory attention networks for skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2021.
- [130] S. Li, W. Li, C. Cook, Y. Gao, and C. Zhu, "Deep independently recurrent neural network (INDRNN)," 2019, *arXiv:1910.06251*.



Tasweer Ahmad received the bachelor's degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2007, and the master's degree in electronics engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2009. He is currently working toward the Ph.D. degree with South China University of Technology, Guangzhou, China.



He has been an Instructor with Government College University, Lahore, Pakistan, from 2010 to 2015, and with COMSATS University Islamabad, Sahiwal Campus, Islamabad, Pakistan, from 2015 to 2016. His current research interests include image processing, computer vision, and machine learning.

Lianwen Jin (Member, IEEE) received the B.S. degree in electronics engineering from the University of Science and Technology of China, Anhui, China, in 1991, and the Ph.D. degree in information and communication engineering from South China University of Technology, Guangzhou, China, in 1996.

He is currently a Professor with the College of Electronic and Information Engineering, South China University of Technology. He has authored more than 100 scientific papers. His research interests include handwriting analysis and recognition, image processing, machine learning, and intelligent systems.

Dr. Jin was the recipient of the New Century Excellent Talent Program of MOE Award and the Guangdong Pearl River Distinguished Professor Award. He is a member of the IEEE Computational Intelligence Society, IEEE Signal Processing Society, and IEEE Computer Society.



Xin Zhang (Member, IEEE) received the B.S. degree in automatic engineering from Northwestern Polytechnical University, Xi'an, China, in 2003, and the M.S. and Ph.D. degree in electrical engineering from Oklahoma State University, Stillwater, OK, USA, in 2005 and 2011, respectively.

She has been a Visiting Scholar with the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, from 2018 to 2020. He is currently an Associate Professor with the School of Electronic and Information Engineering, South China University

of Technology (SCUT), Guangzhou, China. He has authored or coauthored more than 30 articles in journals, books, and conferences, including IEEE Transactions, AAAI, MICCAI, and ACM MM. Her research interests include computer vision, machine learning and medical image analysis.

Dr. Zhang has been the Principal Investigator of several research projects funded by NSFC Foundation, Ministry of Education, Guangdong NSF, and Microsoft Research Asia. She was the reviewer for many international conferences and journals.



Songxuan Lai received the B.S. degree in electronics and information engineering in 2016 from the South China University of Technology, Guangzhou, China, where he is currently working toward the Ph.D. degree in information and communication engineering.

He has authored several publications in high-impact conferences and journals (e.g., ICDAR, AAAI, IEEE TIFS, etc). His research interests include machine learning, OCR systems, handwriting analysis and recognition, and signature verification.



Guozhi Tang received the bachelor's degree in information technology from Yunnan University, Kunming, China, in 2019. He is currently working toward the master's degree in information and communication engineering with South China University of Technology, Guangzhou, China.

His research areas focus on deep learning and car license plate detection.



LuoJun Lin received the B.S. degree in electronic and information engineering from Yunnan University, Kunming, China, in 2014, and the Ph.D. degree in information and communication engineering from the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, in 2020.

She is currently a Lecturer with Fuzhou University, Fuzhou, China. Her research interests include deep learning and computer vision, unsupervised learning, and transfer learning.