Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/50/20241581

An overview of Graph Convolutional Networks in skeleton-based action recognition

Jin Yan

School of Information Science and Engineering, Donghua University, Shanghai, 201620, China

yanjin0726@163.com

Abstract. The field of research is currently focused on human activity recognition. Hence, numerous pertinent literature reviews have expounded upon the multifaceted nature of data, the process of selecting feature vectors, and the advantages and disadvantages of classification networks. Graph Convolutional Networks (GCNs) have demonstrated significant efficacy in the domain of human action recognition. In recent years, with the rapid development of 3D skeleton data collection, a plethora of studies in action recognition based on skeleton data have emerged. Skeleton data consists of three-dimensional coordinates of multiple spatiotemporal skeletal joints, making it an effective representation of kinematics. It can be easily acquired through low-cost depth sensors and also directly extracted from two-dimensional images using video-based pose estimation algorithms, attracting widespread attention. As relational networks continue to evolve, GCNs have been applied to various fields, including human action recognition. GCNs have demonstrated significant advantages in feature extraction from skeleton data. However, using GCNs alone may have various limitations. Therefore, in recent years, many enhancement measures for GCNs have emerged. This review aims to summarize the research achievements of Graph Convolutional Network improvements in the field of human action recognition in recent years. It intends to assist future researchers in quickly organizing their research ideas, facilitating the generation of new results.

Keywords: Human action recognition, Graph Convolutional Networks, enhancement measures

1. Introduction

Human action recognition can be classified into various data representations of frame information, such as RGB video-based action recognition, depth map-based action recognition, and skeletal data-based action recognition. This study centers on the investigation of human action recognition utilizing skeleton data. There are generally two approaches for acquiring skeletal data. One way entails the direct acquisition of spatial coordinates of skeletal joints using wearable hardware devices, whilst the alternative approach involves the utilization of pose estimation algorithms to derive joint coordinates from unprocessed RGB films. The utilization of skeleton data as a type of input for human action recognition commenced at a comparatively later stage when compared to other data input formats, such as RGB films and depth information. During the initial phases of this methodology, the majority of efforts were centered around the utilization of manually crafted features for the purpose of action recognition. One illustrative instance is the work conducted by Vemulapalli et al., wherein they

DOI: 10.54254/2755-2721/50/20241581

conceptualize various anatomical components as sets and establish a mathematical framework to analyze the interconnections between these sets using Lie group algebra [1]. In their study, Yang et al. proposed a multi-task learning approach that specifically targeted the relationship between skeletal components and category labels in order to acquire skeletal feature descriptions [2]. The manually crafted features, while offering a high level of interpretability, have limitations in their ability to capture complex and nonlinear relationships. They are also unable to capture the deeper connections that exist between key points. The RNN series of models possess the ability to preserve correlations between states over a sequence of time steps. The enhanced Long Short-Term Memory network (LSTM) effectively tackles the issue of vanishing gradients in RNNs. Additionally, the Gated Recurrent Unit (GRU) offers a simplified version of the LSTM model. The utilization of RNN-based networks for the purpose of extracting features from sequential data is a widely employed method. However, the suitability of RNN as the ideal solution for tasks of a similar kind is contingent upon the specific attributes of feature vectors at each individual time step.

Convolutional Neural Networks (CNN) are commonly employed in two manners within the context of skeleton-based action recognition. There are two approaches that can be employed in this context. The first approach involves the conversion of skeleton data into pseudo-image data, which can then be processed by a CNN. The second method extracts temporal information using a 1D CNN. In RGB color space, Dennis et al. rebuilt joint positions [3]. This included representing joint three-dimensional spatial coordinates with RGB color channels. The reconstructed image represents joint locations as pixels, and a CNN uses this pseudo-image to recognize activities. CNN and recurrent neural network (RNN) architectures are often used to process grid-structured data with Euclidean spaces properties. Examples include vectorized words in one-dimension, two-dimensional visuals, and three-dimensional videos. This dataset has consistent local context and relative positional consistency, resisting translation. Transforming skeleton data into pseudo-images reframes the difficulty as a deep learning problem in Euclidean space, allowing mechanical extraction of deep features. However, it ignores the spatiotemporal linkages between skeletal joints and their organization in the skeletal structure. Skeleton data is structured data that defies Euclidean geometry. Yan et al. pioneered skeleton data conceptualization as spatiotemporal graph structures and GCN feature extraction [4]. This work was a milestone in related research. Numerous studies have used GCN to develop more systematic and effective feature extraction methods for skeleton data. Today, GCN additions are often the best way to handle datasets.

This paper presents a comprehensive overview of research methodologies employed in diverse improvement methods. It highlights the merits associated with each approach and serves as a valuable reference for researchers in this domain, aiding in the systematic organization of ideas and fostering the progress and refinement of existing methods. It possesses considerable significance for the future advancement in this field.

2. Review of Achievements in Improving Graph Convolutional Networks in the Field of Human Action Recognition

The research on the utilization of GCN for the purpose of extracting features from skeleton data has gained significant attention in recent years. A multitude of scholars have put forth a wide range of enhancement techniques, thereby advancing the utilization of GCN for the extraction of features from skeleton data. Consequently, the subsequent part provides a concise overview of notable improvement methodologies that have emerged in recent years. This serves as a valuable resource for academics in the respective field.

2.1. ST-GCN

Yan et al. presented an innovative end-to-end training technique, representing the initial utilization of GCN in the field of skeleton-based action recognition. A spatiotemporal graph was created, consisting of interconnected nodes and edges. In the context of human anatomy, nodes are utilized to symbolize the various joints present in the body. On the other hand, edges are classified into two distinct categories:

DOI: 10.54254/2755-2721/50/20241581

spatial edges and temporal edges. Spatial edges establish connections between neighboring joints that are physically interconnected within the same frame, whereas temporal edges establish connections between matching joints in consecutive frames. The ST-GCN model makes use of the OpenPose tool in order to extract human key points. The procedure entails dividing the input video into several frames, doing keypoint detection on each frame, and subsequently organizing these key points for subsequent analysis. Figure 1 represents the main model of ST-GCN. The Input Video is formed into a graph based on human keypoints and serves as the input data. The middle part, ST-GCNs, is responsible for feature extraction, and the Readout is used for softmax classification.

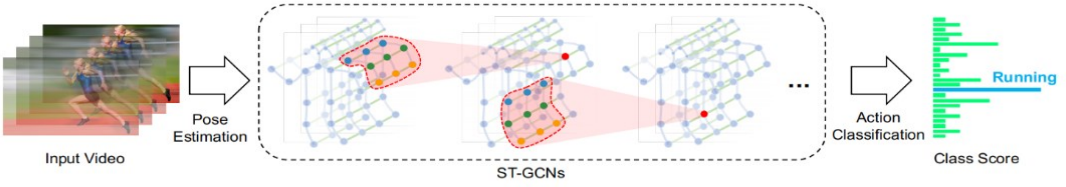

Figure 1. Action Recognition Workflow Diagram [4]

2.2. 2s-AGCN

In their study, Shi et al. expanded the application of ST-GCN for action recognition with the incorporation of adaptive graph convolution layers and second-order information [5]. The authors introduced a novel neural network architecture known as the "two-stream adaptive graph convolutional network" (2s-AGCN). In order to enhance the adaptability of the graph structure, the researchers incorporated a self-attention mechanism and a freely learnt mask. This modification resulted in a graph that is distinct for various layers and samples. In essence, this design imbues distinctiveness to the topological structure of various operations, so facilitating their recognition. Concurrently, the researchers incorporated primary data (joint coordinates) alongside secondary characteristics (body skeleton length and orientation) in order to establish a two-stream framework, as depicted in Figure 2. This network utilizes both joint information and skeletal information through the implementation of two separate streams. The utilization of the two-stream network facilitates the acquisition of a wider range of action features, while employing a score-level fusion technique to amalgamate the outcomes obtained from the two streams.

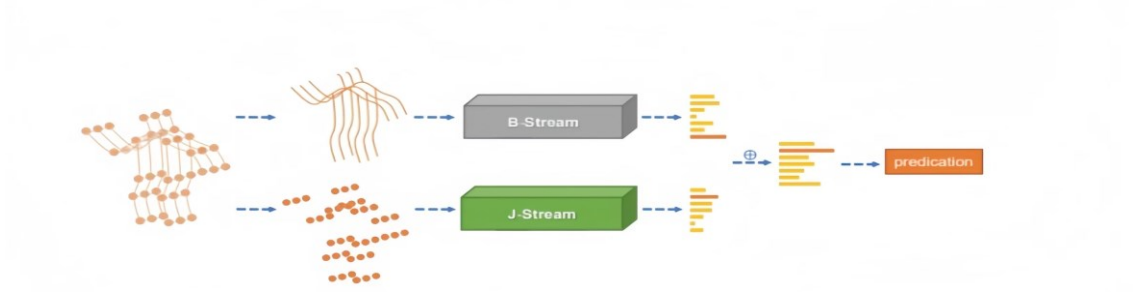

Figure 2. Two-Stream Network [5]

2.3. Shift-GCN

Cheng et al. introduced a novel "Shift Graph Convolutional Network" (Shift-GCN) to address the high computational complexity and lack of flexibility in the receptive fields of spatial and temporal graphs, which are limitations of the GCN method [6]. Shift-GCN combines the "shift" structure from CNN and introduces it into GCN for the first time. It also applies the CNN's shift operation on the temporal Temporal Convolutional Network (TCN), significantly reducing model parameters and computational complexity. In the spatial domain, non-local Shift Graph Convolution is proposed, eliminating the limitations of physically connected nodes and transforming a single-frame skeleton graph into a

DOI: 10.54254/2755-2721/50/20241581

complete graph. Consequently, each node exhibits a direct relationship with every other node. This study presents the introduction of an adaptive non-local transfer mechanism for the purpose of extracting essential human key information from the whole skeleton graph. This mechanism places emphasis on identifying and highlighting significant links within the graph. The authors of this study propose the introduction of an adaptable temporal Shift Graph Convolution in the temporal domain. In the proposed approach, a trainable time offset parameter is learned for each channel. To enhance gradient propagation, the parameters transition from discrete integers to continuous real numbers. This transition is achieved by employing linear interpolation to generate non-integer shifts.

2.4. VE-GCN

In their study, Liu et al. proposed the "Vertex-Edge Graph Convolutional Network" (VE-GCN) as a method for conducting graph convolution operations on selected regions that consist of vertices particular to joints and edges specific to bones [7]. The joints in the human body can be considered as the vertices of a graph, with the bones serving as the edges connecting these vertices. By examining the joint-joint adjacency matrix and the joint-bone adjacency matrix, which are abstract representations of the connections between joints and bones, we can gain insights into the links between distant joints and bones. This methodology integrates characteristics derived from articulations, skeletal structures, and their interconnections. Regarding static features, this study employs joint coordinates and bone directions extracted from each frame. Dynamic characteristics are utilized by employing the temporal displacement of joints and bones between two adjacent frames. In order to achieve a successful integration of static and dynamic elements, researchers employ a two-stream framework. The proposed approach enhances performance by integrating predictions from both the static feature stream and the dynamic feature stream at the score level. In addition, the study incorporates a Conditional Random Field (CRF) model as a loss function to effectively reflect the inherent structural relationships present within the graph outputs.

2.5. TS-GCN

Ding et al. introduced a novel "Temporal Segment Graph Convolutional Networks" (TS-GCN) [8]. The TS-GCN model employs a segmentation technique to divide action sequences into many stages. This approach enables the independent analysis of motion dynamics within each stage, thereby mitigating temporal misalignment issues and enhancing the modeling of action sequences. Following this, an ensemble methodology is employed to amalgamate the outputs from each stage in order to derive the ultimate forecast outcome. To enhance the precision of capturing the temporal dynamics of the complete sequence, an adaptive graph construction module was presented in this study. The module possesses the capability to autonomously acquire and modify the topological configuration of the skeletal graph in response to the varying dynamics seen across distinct stages, hence accommodating diverse skeletal sequences. The integration of a multi-stream framework, which incorporates both rigid body information and joint displacement information alongside the original joint information, leads to enhanced performance in action recognition.

2.6. FGCN

Yang et al. proposed a novel method called "Feedback Graph Convolutional Network" (FGCN) to improve the performance of skeleton-based action recognition [9]. The proposed methodology improves the accuracy and robustness of action recognition by the incorporation of multi-stage temporal sampling algorithms and feedback blocks, which enable the capture of both temporal and contextual information. The FGCN framework employs spatiotemporal information derived from skeleton sequences in order to extract action features. Additionally, it increases the representation of these aspects through the utilization of a feedback mechanism. The researchers devised a multi-stage temporal sampling approach, wherein the input skeleton sequence was divided into many stages. From each stage, a specific number of frames were randomly selected and utilized as input. They introduced a feedback block (FGCB) to transmit information between different stages and enhance feature representation. FGCB achieves

DOI: 10.54254/2755-2721/50/20241581

feature feedback and updates by fusing the features of the current stage with those of the previous stage, using the fused features as the input for the current stage. Furthermore, the researchers compared different spatio-temporal fusion strategies, including "last-win-all fusion","average fusion" and "weighting fusion" with the models fed joints only. The "average fusion" strategy performed the best in this experiment.

2.7. CD-JBF-GCN

Tu et al. proposed a method called "Correlation-driven Joint-Bone Fusion Graph Convolutional Network" for semi-supervised skeleton action recognition [10]. This methodology basically encompasses two distinct procedures: the collaborative process and the osseous process. In the collaborative procedure, the human body's joints function as nodes inside a network, while the bones serve as edges, so establishing a series of skeletal graphs. The primary attributes of nodes consist of the three-dimensional coordinates of the bodily joints they belong to. The bone process involves the conceptualization of human body bones as nodes within a graph, with the joints acting as the edges connecting these nodes. The initial features are derived by subtracting the target joint coordinates from the source joint coordinates, thereby quantifying the spatial disparity between joints as skeletal elements. In order to tackle the matter of information transmission between joints and bones, a module known as CD-JBF-GC is proposed to integrate the motion information of bones into the joint procedure. This module employs correlation matrices and feature transfer functions to facilitate the propagation of information. The correlation matrices describe the connection structure between joints and bones, while the feature transfer functions determine the way features are fused. The researchers experimented with various implementations of feature transfer functions and optimized the feature transfer's effectiveness using trainable weights. By combining the joint and bone processes, the researchers successfully addressed the problem of semi-supervised skeleton action recognition. This method effectively leverages the motion information between joints and bones, improving the accuracy and robustness of action recognition.

2.8. FR-GCN

Huang et al. introduced a novel "Feature Reconstruction Graph Convolutional Network" (FR-GCN) [11]. This study aims to address challenges in skeleton-based action recognition, including the limitations of handcrafted features, incomplete representation of intrinsic connections between joints by deep learning methods, and the need for improvement in the ability and efficiency of existing methods to learn advanced spatio-temporal features. To overcome these challenges, the study proposed a new Graph Convolution Network (GCN) model called FR-GC. This model introduces three key modules: the TPE module (which adaptively divides the original topology and adds new partitions to predefined partitions), the FR-GC module (which enables the interaction of spatial and temporal information, facilitating direct spatio-temporal information flow), and the MS-DTC module (which extends joint sequences to frame-external data and utilizes a sliding time window for cross-frame learning). These modules enhance the ability to extract critical information from skeleton data.

3. Discussion

The application of graph convolutional networks to skeleton-based action recognition was first introduced by ST-GCN. The 2s-AGCN algorithm is a notable advancement built upon the ST-GCN framework, which is widely recognized in the area. Subsequent methodologies have further developed upon the aforementioned algorithms, effectively mitigating their limitations and attaining enhanced performance in feature extraction. Nevertheless, the current methodologies continue to possess some constraints. There exists potential for enhancement. An instance of this can be observed in the Shift-GCN framework, where the introduction of the shift module resulted in a notable decrease in computational complexity. However, it is important to note that this reduction in complexity may potentially result in the loss of information. In the realm of completely supervised learning, the performance of CD-JBF-GCN is rather subpar. Additionally, the successful exploration of the

DOI: 10.54254/2755-2721/50/20241581

association between joints and bones remains an unresolved matter. The parameter count of the model is increased while incorporating the TPE and MS-DTC modules into FR-GCN. Although the incorporation of these modules can improve the performance of the model, it also introduces increased complexity and computational burden. These strategies for improvement serve to boost a certain part of the model's performance, but they may also introduce constraints in other domains. Achieving equilibrium in performance across various aspects poses a significant challenge for future research endeavors.

4. Conclusion

This study presents a comprehensive overview of the use of several graph convolutional networks for the purpose of extracting skeletal data. The objective is to expedite the comprehension of the current research landscape in this domain, hence aiding future researchers in their investigations. Nevertheless, it is important to acknowledge that this article provides a concise overview of a limited selection of prominent methodologies employed in recent years, and may not encompass the entirety of available approaches. Furthermore, a notable deficiency exists in terms of experimental comparisons conducted to evaluate the efficacy of the aforementioned approaches on a shared dataset. Additional investigation is necessary in this domain to mitigate these constraints and offer a more exhaustive analysis of the discipline. Subsequent investigations have the potential to expand upon the aforementioned elements, hence facilitating more enhancements and the optimization of methodologies. Graph convolutional network architectures, model combination strategies, and integration with other deep learning methods may improve skeleton action recognition. Researchers can advance this field by innovating and experimenting with new methods to expand knowledge.

References

[1] Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3d skeletons as points in a lie group[C], in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 588-595

[2] Yang Y, Deng C, Gao S, et al. Discriminative multi-instance multitask learning for 3d action recognition[J]. in: IEEE Transactions on Multimedia, 2016, 19(3): 519-529.

[3] Ludl D, Gulde T, Curio C. Simple yet efficient real-time pose-based action recognition[C]. in: IEEE Intelligent Transportation Systems Conference (ITSC), 2019: 581-588

[4] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition. in: AAAI, 2018: 7444–7452.

[5] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition. in: CVPR, 2019: 1–10.

[6] K. Cheng, Y. Zhang, et al., Skeleton-based action recognition with shift graph convolutional network, in: CVPR, 2020: 1–10.

[7] K. Liu, L. Gao, N.M. Khan, L. Qi, L. Guan,Integrating vertex and edge features with graph convolutional networks for skeleton-based action recognition. in: Neurocomputing, vol. 466, 2021: 190-201

[8] C. Ding, S. Wen, W. Ding, K. Liu, and E. Belyaev, Temporal segment graph convolutional networks for skeleton-based action recognition. in: Engineering Applications of Artificial Intelligence, vol. 110, 2022: 104675,.

[9] H. Yang, D. Yan, L. Zhang, Y. Sun, D. Li and S. J. Maybank, Feedback Graph Convolutional Network for Skeleton-Based Action Recognition. in: IEEE transactions on image processing, 2022: 164-175.

[10] Z. Tu, J. Zhang, H. Li, Y. Chen and J. Yuan, Joint-Bone Fusion Graph Convolutional Network for Semi-Supervised Skeleton Action Recognition. in: IEEE Transactions on Multimedia, vol. 25, 2023: 1819-1831.

DOI: 10.54254/2755-2721/50/20241581

[11] J. Huang, Z. Wang, J. Peng, F. Huang, Feature reconstruction graph convolutional network for skeleton-based action recognition. in: Engineering Applications of Artificial Intelligence, vol. 126, Part B, 2023: 106855