EURASIP Journal on Advances
in Signal Processing

RESEARCH **Open Access**

Enhancing human behavior recognition with spatiotemporal graph convolutional neural networks and skeleton sequences

Jianmin Xu[1], Fenglin Liu[2], Qinghui Wang[2], Ruirui Zou[2], Ying Wang[2], Junling Zheng[3], Shaoyi Du[4] and Wei Zeng[2*]

*Correspondence:
zengwei@lyun.edu.cn

[1] School of Sports and Health, Longyan University, Longyan 364012, China
[2] School of Physics and Mechanical and Electrical Engineering, Longyan University, Longyan 364012, China
[3] School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[4] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China

Abstract

Objectives: This study aims to enhance supervised human activity recognition based on spatiotemporal graph convolutional neural networks by addressing two key challenges: (1) extracting local spatial feature information from implicit joint connections that is unobtainable through standard graph convolutions on natural joint connections alone. (2) Capturing long-range temporal dependencies that extend beyond the limited temporal receptive fields of conventional temporal convolutions.

Methods: To achieve these objectives, we propose three novel modules integrated into the spatiotemporal graph convolutional framework: (1) a connectivity feature extraction module that employs attention to model implicit joint connections and extract their local spatial features. (2) A long-range frame difference feature extraction module that captures extensive temporal context by considering larger frame intervals. (3) A coordinate transformation module that enhances spatial representation by fusing Cartesian and spherical coordinate systems.

Findings: Evaluation across multiple datasets demonstrates that the proposed method achieves significant improvements over baseline networks, with the highest accuracy gains of 2.76% on the NTU-RGB+D 60 dataset (Cross-subject), 4.1% on NTU-RGB+D 120 (Cross-subject), and 4.3% on Kinetics (Top-1), outperforming current state-of-the-art algorithms. This paper delves into the realm of behavior recognition technology, a cornerstone of autonomous systems, and presents a novel approach that enhances the accuracy and precision of human activity recognition.

Keywords: Behavior recognition, Graph neural networks, Skeleton sequences, Coordinate transformation, Connection feature

1 Introduction

The task of human behavior recognition aims to enable computers to identify the current action category of the subject based on data captured by cameras, radar, or other sensors. This research task is highly challenging, demanding, and valuable in terms of application. Human motion data are primarily obtained from cameras and various types of sensors, leading to two main data formats. One format involves raw RGB image data

[1], which is primarily acquired through traditional image sensors, i.e., cameras. The other format involves human skeletal point data, composed of two-dimensional or three-dimensional coordinates of key joints in the human body. This skeletal point data can be obtained through specific sensors or by processing depth-enhanced raw RGB image data using pose estimation algorithms. These give rise to two distinct recognition methods: those utilizing RGB data and those utilizing skeletal point data [2]. Each of these methods comes with its own advantages and limitations. Recognition methods based on RGB images retain complete semantic information from the original images, yielding high recognition accuracy; however, they consume substantial computational resources and involve complex algorithms. On the other hand, methods employing skeletal point data are advantageous due to their convenience in processing, low computational resource requirements, and reduced algorithm complexity. Nevertheless, a drawback is the potential loss of some semantic information during the acquisition of skeletal point data, leading to slightly lower recognition accuracy compared to RGB-based methods. This paper primarily explores the utilization of skeletal point data, which holds promising application prospects in various fields.

First and foremost, the application in production and daily life environments stands out. With the rapid development in the field of artificial intelligence, intelligent robots have become prevalent in people's lives. The trend of using machines to replace human labor is becoming more pronounced. Traditional machines often require human intervention, with operators manipulating them from control stations. Incorporating human behavior recognition technology into intelligent robots enables a decoupling between the robots and operators. Robots can comprehend task instructions through gestures and movements of the operators, thereby reducing the burden on human workers. For tasks that necessitate human collaboration, the application of human behavior recognition technology makes human–machine interaction more convenient and enhances work efficiency. Secondly, the application in intelligent surveillance is noteworthy. Video surveillance is a common tool in both industrial and daily life contexts, ranging from homes and streets to hospitals and factories. However, the processing of the video footage generated by these surveillance systems can be challenging. Manual observation is time-consuming and inefficient. Human behavior recognition technology offers an effective solution. By integrating this technology, abnormal human behaviors in surveillance videos can be rapidly identified. In domestic environments, intelligent surveillance can quickly detect intruders and raise alarms promptly. For households with elderly individuals, it can swiftly detect anomalies in their conditions, such as falls or unconsciousness, and notify family members or authorities, thereby reducing potential harm to both individuals and property. In criminal investigations, it aids law enforcement in swiftly locating suspects. In industrial settings, it can detect improper operations and issue timely warnings to prevent accidents. Human behavior recognition technology also finds application in the entertainment sector, including virtual reality technologies and VR games. Additionally, due to the rapid growth of short video platforms such as TikTok and Kwai, enormous amounts of video data are uploaded daily, posing significant challenges in content moderation. Human behavior recognition technology can efficiently alleviate the workload of content reviewers by rapidly identifying whether the actions in uploaded videos violate guidelines. Currently, many applications of human behavior recognition technology are still in their infancy. Nevertheless, with the maturation of technology and advancements in hardware, the scope of application for this technology is expected to expand extensively.

Human behavior recognition technology finds wide application across various domains. Human skeletal point data have garnered attention due to its low computational cost and minimal susceptibility to environmental interference. The process of utilizing skeletal point data for behavior recognition encompasses three main steps. The first step is action acquisition, generally facilitated by image sensors and depth sensors. Subsequently, skeletal point data generation follows. This entails employing pose estimation algorithms to extract key human joint positions from the acquired images. These joint positions are then sequenced over time, resulting in organized skeletal point data. The final step involves behavior recognition. This entails employing behavior recognition algorithms to extract features from the skeletal point data. These extracted features are then input to a classifier to determine the final action category.

There are generally four types of human behavior recognition algorithms: traditional methods, recurrent neural networks, convolutional neural networks, and graph convolutional neural networks. Traditional methods primarily utilize manual feature engineering for behavior recognition. These methods include approaches like those by Laptev [3], who use interest point detection to map 3D videos into 1D space, effectively identifying moving objects. Dalal et al. [4] proposed the use of histogram of oriented gradients (HOG) to detect human outlines, leveraging gradient information to compute appearance features. Histogram of Flow (HoF) descriptors exploit optical flow information to compute temporal transformations and inter-frame correlations. Oreifej et al. [5] introduced a novel action recognition descriptor called the HON4D descriptor, which combined depth image sequences, spatial positions, and temporal information to map features into a four-dimensional space. Vemulapalli et al. [6] used various skeletal point description techniques to represent actions, mapping action curves into the Lie algebra vector space.

Recurrent neural networks (RNNs) are widely utilized for behavior recognition due to their distinctive advantage in handling time-series data. Among these methods, long short-term memory (LSTM) networks are the most extensively employed. Du et al. [7] proposed an end-to-end hierarchical RNN algorithm, wherein the human skeletal representation is divided into five segments. LSTM is applied to process each of these segments separately, extracting temporal information from each part. The extracted information is then progressively fused across these segments, resulting in comprehensive spatial feature information. A classifier is employed to generate a probability matrix. Song et al. [8] introduced STA-LSTM, which incorporates spatiotemporal attention mechanisms into the LSTM main network. The main network utilizes LSTM to extract features, and spatiotemporal attention weights are calculated before and after the main network to incorporate temporal and spatial aspects. Zhang et al. [9] introduced attention gate structures within each RNN neuron, enabling each neuron to generate distinct attention effects for various inputs. Qiu et al. [10] introduced the concept of spatiotemporal saliency and integrated it into LSTM to enhance the significance of spatiotemporal features. Additionally, a dual-stream fusion approach is adopted to improve recognition accuracy.

Behavior recognition methods based on convolutional neural networks (CNNs) employ convolutional operations to extract spatiotemporal feature information. For instance, in TS-LSTM [11], the authors employ convolutional layers to extract intra-segment spatial feature information. Subsequently, either CNN or RNN is utilized to capture temporal features. The temporal Conv [12] method employs temporal convolutional neural networks to extract temporal features, serving as an alternative to long short-term memory networks. In the HCN method [13], a hierarchical convolutional network is proposed to learn co-occurrence features. This approach aims to aggregate contextual information across different hierarchical levels. Through this layered convolutional network, the method seeks to harness context information at various levels effectively.

Graph convolutional networks (GCNs) are highly suitable for handling non-Euclidean distance data, making them particularly well-suited for processing skeletal point data. Spatiotemporal graph convolutional network (ST-GCN) [14] was among the first to extend graph convolutions to skeletal point data. It performs feature extraction on single-frame skeletal point data using a graph convolutional network. By utilizing human joint connections as the graph structure, it effectively extracts spatial feature information. Subsequently, a one-dimensional convolution operation processes the spatial feature information extracted from each frame in the video sequence to obtain temporal features. A classifier is then employed to calculate class probabilities. Building upon ST-GCN, several improved methods have been introduced. In AS-GCN [15], the authors enhance the consideration of dependencies among joints that are not physically adjacent in the human body. They introduce action connections and structure connections to capture dependencies between any nodes. In 2 s-AGCN [16], adaptive graph convolution operations and a dual-stream structure are employed to enhance recognition accuracy. Shift-GCN [17] reduces computational complexity by replacing convolutions with channel shifting operations. It introduces adaptive spatial shifting and temporal shifting operations to enhance spatiotemporal features. The MS-G3D algorithm [18] introduces a concept of multi-scale decoupling. Recognizing that dependencies between nodes in the structural graph of graph convolutions are overly coupled and lack temporal structure, MS-G3D introduces the G3D operator. This operator aims to decouple joint spatial dependencies and introduce time dependencies. The methods mentioned above are all supervised approaches. However, supervised methods have limited applicability and can only be used in scenarios where sufficient labeled data is available. For scenarios with insufficient labeled data, supervised methods may not perform well. As a result, researchers have turned their attention to studying unsupervised methods for human behavior recognition using skeletal point data. The authors of LongT GAN [19] designed a skeletal point sequence restoration architecture to learn fixed-dimensional representations. They employed additional adversarial training strategies as guidance. MSL researchers [20] found that learning feature representations solely from a single reconstruction task could lead to overfitting and inadequate feature representations for behavior recognition. They proposed a combined approach involving multiple tasks, including behavior prediction, jigsaw puzzles, and contrastive learning, to learn diverse skeletal point features. Motion feature information can be modeled through behavior prediction, jigsaw puzzles are used for learning temporal features, and contrastive learning regulates the features extracted from different tasks. AS-CAL [21] introduced a contrastive action learning paradigm to enhance action patterns in skeletal point sequences. They proposed a momentum LSTM as an encoder and a novel action representation called Contrastive Action Encoding (CAE). PC-Net [22] utilized a Gated Recurrent Unit (GRU) as an encoder and decoder. By fixing states and weights, the decoder's influence was weakened, forcing the encoder to learn features that better represent action categories. SeBiReNet [23] introduced a tethered denoising autoencoder for learning 3D pose representations. It can separate pose-related and view-related features from skeletal data entirely in an unsupervised manner. The authors also proposed a sequential bidirectional recursive network to model skeletal data. CrosSCLR [24] first introduced a highly consistent skeletal point contrastive representation, which can capture contrastive learning samples with high similarity. To address the limitation of information extracted from a single view representation, they proposed cross-view feature exchange to enhance the accuracy of feature extraction.

This paper presents a supervised human behavior recognition method based on spatiotemporal graph convolutional neural networks, referred to as the coordinate transformation and connectivity feature-based human behavior recognition method. When processing sequences of human skeletal point data, this method effectively extracts joint connectivity feature information, enhancing spatial feature representation, while also proficiently capturing temporal variations. Compared to the baseline method, it demonstrates superior recognition accuracy.

Graph convolutional neural networks have yielded impressive recognition outcomes in behavior recognition based on skeletal data. The graph structure, a fundamental component of graph convolutional neural networks, encapsulates relationships among nodes within non-Euclidean data. Concealed within this graph structure are connectivity features between nodes that can furnish supplementary spatial features denoting inter-joint relationships. However, numerous methods employing graph convolutional neural networks overlook these spatial features. This paper introduces a connectivity feature extraction module to acquire implicit connections between human joints, extracting hidden spatial features from both structural and implicit joint connections. To enhance temporal feature representation, a long-range frame difference feature extraction module is proposed, employing extensive frame differences to achieve a larger temporal receptive field. Additionally, a coordinate transformation module is devised to convert human skeletal points from Cartesian coordinates to spherical coordinates while simultaneously retaining the features of both coordinate systems, thus acquiring more comprehensive features. Finally, through multi-stream fusion, the outputs are combined, leveraging advantages from different perspectives to further elevate recognition accuracy. Experimental results underscore the efficacy of the three proposed modules in enhancing feature representation. The ultimate method exhibits significant improvement over baseline networks and even achieves promising outcomes compared to current leading-edge algorithms across multiple datasets.

2 Problem description

The development of human behavior recognition technology based on skeletal point data has reached an advanced stage, with a substantial portion of credit attributed to the application of spatiotemporal graph convolutional neural networks. These networks have showcased remarkable potential within the realm of behavior recognition, and in recent years, numerous methods have embraced the concepts of spatiotemporal graph convolutions.

The interconnections among human joints constitute a pivotal component of skeletal point-based human behavior recognition methodologies. Many approaches regard these interconnections as the graph structure of graph convolutional neural networks. In the case of ST-GCN [14], the network's graph structure solely incorporates structural connections among human body components, failing to consider potential relationships between non-physically connected joints. Building upon the foundation of ST-GCN, AS-GCN [15] introduced action-specific connections and structural connections, tailoring distinct action connections for different actions and integrating these two types of connections into a novel graph structure. Meanwhile, 2 s-agcn [16] leverages adaptive graph convolution modules to capture implicit connections within the human body. The MS-G3D [18] method employs a decoupled multi-scale aggregation approach to generate multi-scale graph structures. These methods primarily treat the interconnections between human joints as the graph structure of graph convolutional networks, neglecting the spatial features inherently embedded within these interconnections. The extraction of these features warrants further investigation.

Temporal features also hold significance in human behavior recognition. Many techniques employ one-dimensional convolutions to capture temporal dimension features. Due to the constraints of convolutional operations, one-dimensional convolutions can only access features between adjacent frames. Frame difference representation serves as an effective means to capture temporal features. In SGN [25], authors utilize frame differences between adjacent frames as dynamic representation features. In Shift-GCN [17], authors employed frame differences between neighboring frames as inputs for a stream within their network, thereby extracting temporal features. These methods overlook long-term dependencies along the temporal dimension, and extracting long-term temporal dependencies also serves as an avenue for feature enhancement. Human skeletal point data comprises a set of joint coordinate data represented in a three-dimensional Cartesian coordinate system. Concerning the representation of human joints, the relationships among three-dimensional coordinates are sparse, and the connections between joints lack density. Utilizing skeletal point data solely based on three-dimensional Cartesian coordinates results in overly simplistic feature information for human behavior recognition.

This paper will expound upon the aforementioned three issues and elucidate the proposed solutions for these challenges. Through experimental validation, the efficacy of the solutions introduced in this paper is demonstrated.

3 Method

3.1 Overall network architecture

The overall framework of the proposed human behavior recognition algorithm based on coordinate transformation and connectivity features is depicted in Fig. 1. Initially, the input skeletal point sequence data are processed through two channels to extract temporal and spatial feature information. The long-distance frame difference feature extraction module in Fig. 1 is responsible for capturing temporal features. The roles of the

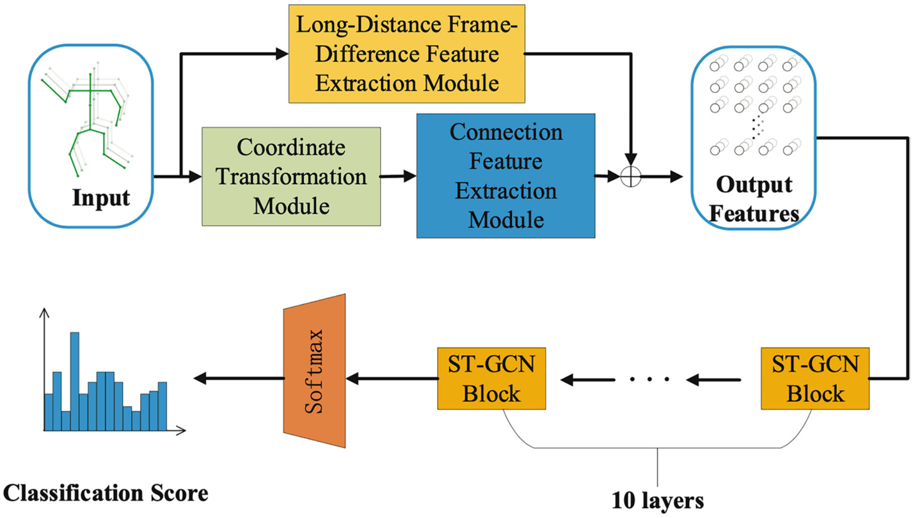

Long-Distance Frame-Difference Feature Extraction Module
Input
Coordinate Transformation Module
Connection Feature Extraction Module
Output Features
Softmax
ST-GCN Block
· · ·
ST-GCN Block
Classification Score
10 layers

Fig. 1 Network architecture diagram

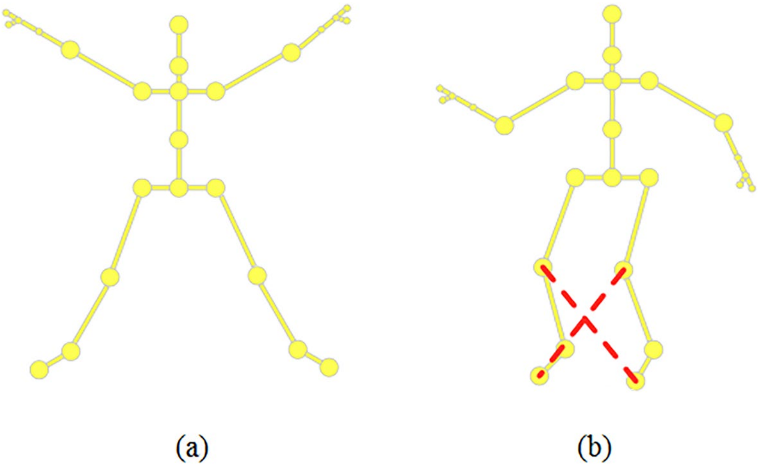


Fig. 2 Natural and implicit connections of human joints: **a** natural connection; and **b** implicit connection

coordinate transformation module and the connectivity feature extraction module are to acquire spatial features. The features extracted from both channels are then combined using summation to obtain subsequent output features. These output features are fed into the subsequent network, where ST-GCN is employed in this study. Finally, the output features are input to a classifier, which produces the predicted class.

3.2 Connectivity feature extraction module

Human joint connections can be categorized into two types: natural connections and implicit connections. Natural connections are formed based on the appearance of the human body, as illustrated in Fig. 2a. The red connections in Fig. 2b represent implicit connections between the leg and knee joints during running. This paper employs an attention module to capture implicit connections among human joints, with the specific calculation process outlined in Eq. (1).

$$A_{hid} = softmax\left(\frac{(W_1X)\cdot(W_2X)^T}{\sqrt{dim}}\right) \tag{1}$$

where X represents the input skeletal point data, W denotes the learnable weight matrix, dim stands for the channel dimension, and A represents the output implicit connectivity matrix.

In the current state, methods based on graph convolutional neural networks consider human body connectivity features as a graph structure, often overlooking the wealth of information contained within these connections. The joints' connections encompass significant local spatial feature information that can enhance the global spatial features extracted by graph convolutional neural networks. This paper introduces the connectivity feature extraction module to extract local spatial features embedded within human joint connections. The network architecture of this module is depicted in Fig. 3.

The attention module in the diagram is utilized to obtain the implicit connectivity matrix of human joints. Once the implicit connections of the human joints are acquired, the connectivity feature extraction module employs two linear layers and an activation function to separately extract the local spatial feature information from the natural and implicit connections of the joints. The obtained features are then added to the original input, enhancing the spatial features of the input. This enhanced input is subsequently utilized as the input for the following network. The specific calculation formula is depicted as follows:

$$\tilde{X} = X + \sigma(W_5(\sigma(W_3A_{norm}) + 0.5 * \sigma(W_4A_{hid}))) \tag{2}$$

where $\tilde{X}$ is the output feature, X is the input feature, σ is the ReLU activation function, A_{norm} is the natural connectivity matrix of human joints, A_{hid} is the implicit connectivity matrix from the attention module, and W_3, W_4, and W_5 are learnable weight matrices.

3.3 Long-distance frame difference feature extraction module

The temporal context information is also crucial for the recognition of human body behaviors based on skeletal point data. Many previous approaches utilized one-dimensional temporal convolution operations in the time domain to extract contextual feature information. The size of the temporal receptive field depends on the size of the convolution kernel. However, constrained by computational resources, convolution kernels are generally small, which leads to the inability of traditional temporal convolutions to

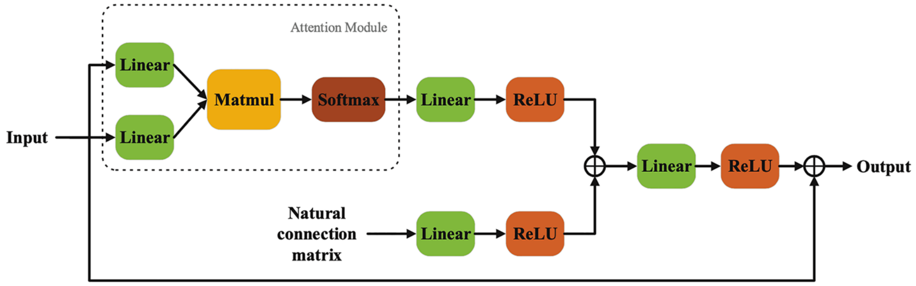


Fig. 3 Network architecture diagram of the connectivity feature extraction module

capture long-range temporal dependencies. This paper proposes the use of long-range frame difference features to obtain distant temporal dependencies, as shown in Fig. 4. By considering larger time intervals between frames, this module can capture more extensive temporal context information, enabling a better understanding of the evolution of actions over time.

The specific calculation formula for extracting long-distance frame difference features, proposed by this method's long-distance frame difference extraction module, for capturing long-range temporal dependencies is given as Eq. (3):

$$D_t = \sigma(W_5\sigma(W_6(x_{t-d} - X_t))) \tag{3}$$

where D_t is the extracted long-range temporal feature at time t, X_t is the skeleton data at time t, x_{t-d} is the skeleton data at time $(t-d)$, d is a hyperparameter representing the inter-frame distance, W_5 and W_6 are learnable weight matrices, and σ is the ReLU activation function.

3.4 Coordinate transformation module

Currently, human body behavior recognition methods based on skeletal point data use the three-dimensional Cartesian coordinates of human joints as inputs to the network, as shown in Fig. 5a. When describing the coordinates of human joints in a three-dimensional Cartesian coordinate system, the human joints are represented in the form of (x_t, y_t, z_t). The three-dimensional Cartesian coordinates focus on the spatial position information of human joints, effectively capturing the pose states of different body parts during human motion. However, using three-dimensional Cartesian coordinate representation for human joint data in behavior recognition has two disadvantages. Firstly, the coordinates represented in the three-dimensional Cartesian system are linear transformations, and changes in all three dimensions can be obtained through translation. However, human motion is more like rotational movement around joints, which cannot be adequately represented in the three-dimensional Cartesian coordinate system. Secondly, the representation of three-dimensional Cartesian data along the axes is not interrelated. Human motion is a holistic action, not separate movements along each axis. Therefore, using three-dimensional Cartesian coordinate representation of human joint data as input to

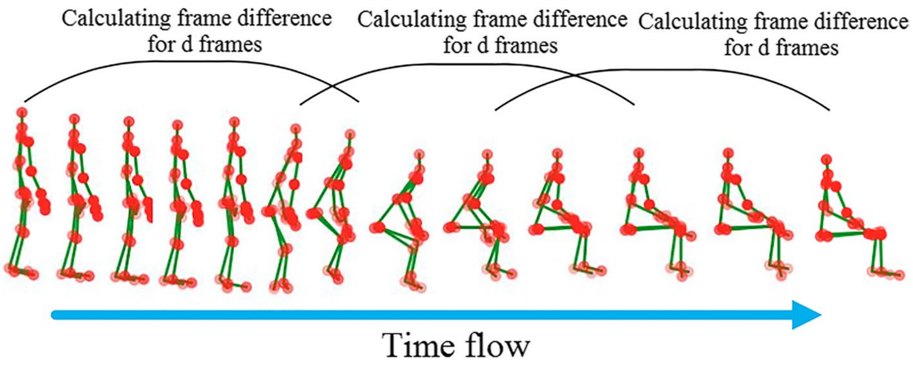

Calculating frame difference for d frames
Calculating frame difference for d frames
Calculating frame difference for d frames
Time flow

Fig. 4 Extraction of long-distance frame differences in human skeleton points

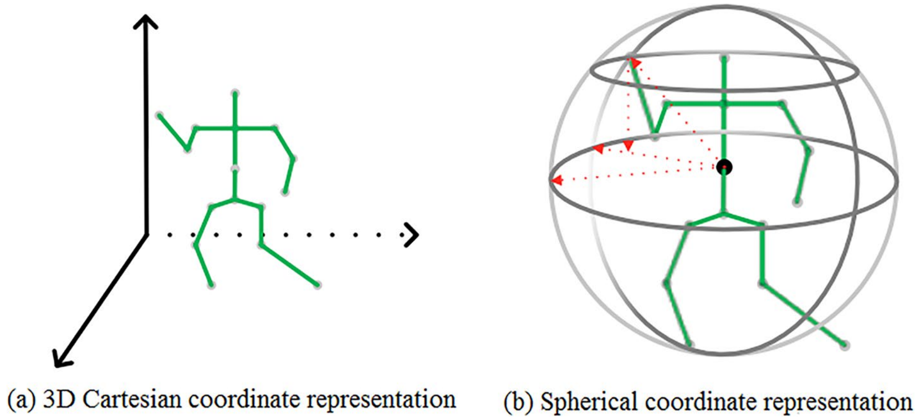

(a) 3D Cartesian coordinate representation (b) Spherical coordinate representation

Fig. 5 Overall framework

the network will undoubtedly lose the interconnected information between different coordinate axes. Representing human joint points using spherical coordinates, as opposed to using Cartesian coordinates, focuses more on the relationship between each joint and the body's center of gravity, as shown in Fig. 5b. Each joint point of the human body can be represented using the distance and rotational angle relative to the body's center of gravity. This representation effectively captures the nonlinear morphological changes during human motion.

The proposed method in this paper utilizes a coordinate transformation module to convert the representation from three-dimensional Cartesian coordinates to spherical coordinates. This transformation is aimed at capturing the nonlinear features carried by human motion. The basis for the coordinate transformation is given by Eqs. (4) to (6). Additionally, to retain the advantages of the three-dimensional Cartesian coordinate representation, this paper suggests using both types of coordinate representations for human skeletal point data as inputs to the network. This approach enables the extraction of both linear and nonlinear features present in human motion.

$$R_{t,k} = \sqrt{\hat{x}^2 + \hat{y}^2 + \hat{z}^2} \tag{4}$$

$$\theta_{t,k} = arccos(\frac{z_{t,k} - z_t}{R_{t,k}}) \tag{5}$$

$$\Psi_{t,k} = arctan(\frac{y_{t,k} - y_t}{x_{t,k} - x_t}) \tag{6}$$

In Eq. (4), $\hat{x} = x_{t,k} - x_t$ represents the difference in x-coordinate between the center of mass of the human body at time t, denoted as (x_t, y_t, z_t), and the three-dimensional Cartesian coordinates of joint k at time t, denoted as $(x_{t,k}, y_{t,k}, z_{t,k})$. Furthermore, $R_{t,k}, \theta_{t,k}, \Psi_{t,k}$ signifies the spherical coordinate representation of joint k of the human body at time t.

3.5 Multi-stream fusion module

The dual-stream network architecture is a classic structure in the field of behavior recognition. It was initially used for behavior recognition based on RGB videos, where it combines RGB and optical flow information. The RGB stream is used to capture pose information, while the optical flow stream captures motion information. By utilizing two streams with different representations, this architecture obtains distinct meaningful features. This approach effectively enhances the accuracy of the algorithm. In the context of the 2 s-AGCN framework, the dual-stream architecture was first applied to human body behavior recognition tasks based on skeletal data, achieving promising results. In a similar vein, this study draws inspiration from its network architecture and incorporates the multi-coordinate representation data proposed in this work as separate streams for network input. For each input stream, a separate training process is conducted, and the output probability matrices from each stream are averaged to obtain a new probability matrix, which serves as the final recognition result. This strategy leverages multiple streams to capture different aspects of the input data, contributing to improved recognition outcomes.

As depicted in Fig. 6, the schematic diagram illustrates the multi-stream fusion approach employed in this chapter. The multi-stream fusion method involves two distinct input streams: the joint data stream, depicted as the "joint" stream in the diagram, and the human skeleton data stream, referred to as the "bone" stream. The joint data stream encompasses the three-dimensional coordinates of human joints, denoted as $(x_{t,k}, y_{t,k}, z_{t,k})$, which represent the coordinate position of joint k at time t. The human skeleton data stream is derived from processed joint data. For two naturally connected joints (x_1, y_1, z_1) and (x_2, y_2, z_2) at the same time instant, the corresponding bone data is calculated as $v_{1,2} = (x_2 - x_1, y_2 - y_1, z_2 - z_1)$. The complete set of naturally connected joints forms the human skeleton data. Since the human joint structure constitutes an acyclic graph, the number of edges is one less than the number of nodes. To address this disparity, self-loop edges are added to the center of mass node, ensuring that both streams have the same data shape. Building upon the joint and skeleton streams, this method maps both streams to spherical coordinate systems to provide additional feature information. This mapping culminates in the multi-stream fusion structure depicted in

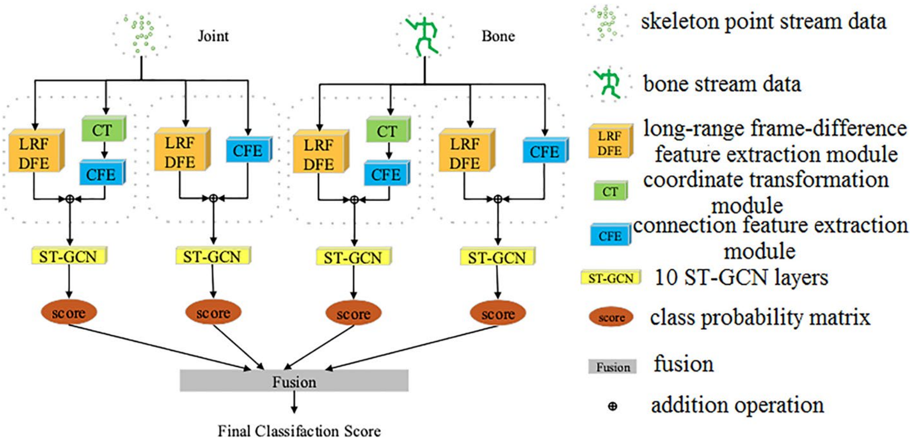

Joint
Bone
LRF DFE
CT
CFE
LRF DFE
CFE
LRF DFE
CT
CFE
LRF DFE
CFE
ST-GCN
ST-GCN
ST-GCN
ST-GCN
score
score
score
score
Fusion
Final Classifaction Score
skeleton point stream data
bone stream data
LRF DFE long-range frame-difference feature extraction module
CT coordinate transformation module
CFE connection feature extraction module
ST-GCN 10 ST-GCN layers
score class probability matrix
Fusion fusion
⊕ addition operation

Fig. 6 Schematic diagram of multi-stream fusion structure

the diagram. By incorporating both joint and skeleton data, and further enhancing their representation in spherical coordinates, this multi-stream fusion framework enhances the model's ability to capture a broader range of features.

4 Experiments and analysis

4.1 Datasets and evaluation metrics

To validate the effectiveness of the proposed method, this study conducted verification experiments on multiple datasets: NTU-RGB+D 60 [26], NTU-RGB+D 120 [27], and Kinetics [28]. Each dataset is associated with its respective evaluation criteria.

First, let us consider the NTU-RGB+D 60 dataset, commonly referred to as the NTU 60 dataset. This dataset comprises a total of 56,880 videos. Each video consists of three types of data: RGB video, IR video, and 3D skeletal point data. For this study, only the skeletal point data was utilized. Within the NTU 60 dataset, there are a total of 60 action classes. The actions were captured using three cameras, and the 3D skeletal point data was captured using Kinect depth sensors. In the NTU-RGB+D skeletal point dataset, each action sequence is represented by 25 skeletal points, as depicted in Fig. 7a. Additionally, only up to two action subjects were considered in each actual action frame. The creators of the NTU 60 dataset recommend the use of two evaluation criteria: cross-view and cross-subject. The cross-view criterion differentiates the training and testing sets based on the sensor ID. Conversely, the cross-subject criterion divides the videos into training and testing sets based on different performing groups. Each group consists of 20 volunteers who performed the actions for the captured videos.

Next is the NTU-RGB+D 120 dataset, abbreviated as NTU 120 dataset. As the name suggests, it is an expanded version of NTU 60 dataset, where the number of action classes has been increased from 60 to 120, and the overall dataset size has roughly doubled. The sample format and human skeletal representation in NTU 120

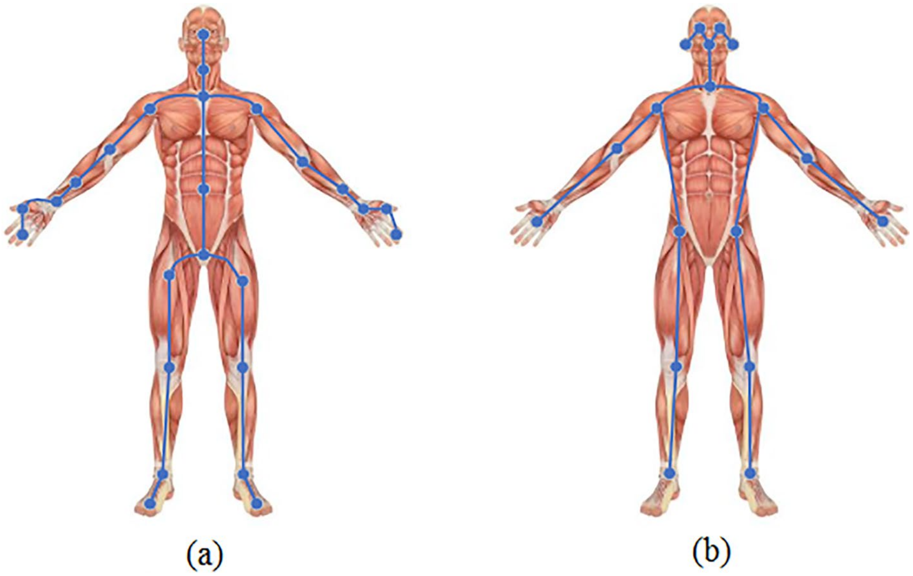


Fig. 7 Schematic diagram of human skeletal structure: **a** NTU-RGB+D human skeletal structure; **b** kinetics human skeletal structure

remain consistent with NTU 60. The key difference lies in the fact that NTU-RGB+D 120 dataset includes a broader range of subjects, expanding from 40 individuals to 106 individuals. Additionally, the sensor setup involves 32 different configurations, requiring variations in working locations and backgrounds for data collection. The cross-setup evaluation criterion divides the 32 sets of data into even and odd groups, with even-numbered sets as training and odd-numbered sets as testing.

Finally, let us discuss the Kinetics dataset. The Kinetics dataset is a large-scale dataset containing over 300,000 video clips with a total of 400 action classes. This dataset comprises only RGB videos and does not include skeletal point data. In this study, pose estimation methods are employed to extract skeletal coordinate information from the RGB videos. Each individual in the extracted action sequences is represented by 18 skeletal points. When multiple individuals appear in a video, only the skeletal data of two individuals are retained. The skeletal representation of an action sequence in the Kinetics dataset is illustrated in Fig. 7b. Unlike the NTU-RGB+D dataset, the Kinetics dataset is not captured using specialized cameras and controlled conditions. Instead, the videos are captured in real-world environments, resulting in significant variations between samples and an uneven distribution of videos. Due to these challenges, training models on the Kinetics dataset can be difficult. Furthermore, the Kinetics dataset lacks professionally captured skeletal point data and relies on pose estimation tools, introducing considerable inaccuracies. As a result, achieving high recognition accuracy for behavior recognition on the Kinetics dataset remains challenging for current algorithms.

The information of the three datasets are summarized as follows.

1. NTU-RGB+D 60:
   - Total number of videos: 56,880.
   - Number of action classes: 60.
2. NTU-RGB+D 120:
   - Total number of videos: Approximately double the size of NTU-RGB+D 60.
   - Number of action classes: 120.
   - Distribution of action classes: Similar to NTU-RGB+D 60, we have provided a table summarizing the number of videos per action class.
3. Kinetics:
   - Total number of video clips: Over 300,000.
   - Number of action classes: 400.
   - Distribution of action classes: We have acknowledged the uneven distribution of videos across action classes in the Kinetics dataset and discussed its potential impact on the interpretation of performance scores.

In this paper, the evaluation criteria proposed by the authors are followed for the NTU-RGB+D dataset, where the maximum probability action is used as the recognition result. For the Kinetics dataset, the evaluation is based on both top-1 and top-5 accuracy metrics.

4.2 Experimental environment and parameters

The implementation framework of the code for this paper's model is based on PyTorch, and the model training and testing are conducted using an NVIDIA GeForce 2080 device. The frame difference span for the long-distance frame difference feature extraction module in the model is set to 5 frames. The datasets used in this paper are aligned by duplicating or reducing frames to facilitate network training. For the NTU-RGB+D dataset, each action sequence consists of 300 frames, while the Kinetics dataset retains 150 frames. Only 2 action subjects are retained for each dataset by selecting the maximum values. During training on the NTU dataset, the initial learning rate is set to 0.05, and it is reduced to one-tenth at the 30th and 40th epochs, for a total of 50 epochs. For training on the Kinetics dataset, the initial learning rate is set to 0.1, and it is reduced to one-tenth at the 40th and 50th epochs, for a total of 65 epochs.

4.3 Ablation experiments

In this subsection, the effectiveness of the proposed method is validated through ablation experiments using the NTU 60 dataset and the cross-subject evaluation criterion. To verify the impact of different coordinate representations, experiments are conducted using two coordinate representations. Table 1 presents the specific experimental results. In the table, "CFE" represents the connection feature extraction module, and "LRFDFE" represents the long-range frame difference feature extraction module. The "baseline" refers to the ST-GCN method, which achieves 86.30% accuracy under the cross-subject evaluation criterion through data preprocessing and parameter tuning.

The primary purpose of Table 1 is to conduct an ablation study, investigating the individual and combined effects of the proposed modules: the CFE module and the LRFDFE module. The baseline accuracy of 86.30% corresponds to the performance of the ST-GCN method without the integration of our proposed modules. This baseline is established to provide a reference point for evaluating the improvements introduced by our contributions. In the table, the rows represent different experimental settings, where we selectively enable or disable the proposed modules to analyze their impact on the overall recognition accuracy. The first row (86.30%) represents the baseline performance without any of our proposed modules. The subsequent rows introduce the CFE module and the LRFDFE module individually and in combination, allowing us to quantify their respective contributions. Specifically, the accuracy of 86.67% corresponds to the

Table 1 Ablation experiments on NTU 60 dataset using cross-subject evaluation

Coordinate representation	CFE module	LRFDFE module	Accuracy (%)
Rectangular coordinates	×	×	86.30
	√	×	86.67
	×	√	87.74
	√	√	88.03
Spherical coordinates	×	×	86.30
	√	×	87.12
	×	√	88.06
	√	√	88.12
Rectangular + spherical coordinates	√	√	89.06

scenario where only the CFE module is enabled, while the LRFDFE module remains disabled. This value highlights the performance improvement achieved solely by incorporating the connection feature extraction module into the baseline method. Similarly, the accuracy of 87.74% corresponds to the scenario where only the LRFDFE module is enabled, while the CFE module is disabled. This value demonstrates the performance gain attributed to the long-range frame difference feature extraction module alone. The final row, with an accuracy of 88.03% (rectangular coordinates) and 88.12% (spherical coordinates), represents the scenario where both the CFE and LRFDFE modules are enabled simultaneously. These values reflect the cumulative effect of combining the two proposed modules, showcasing their complementary contributions to improving the overall recognition performance.

In order to investigate the effectiveness of various modules in this approach, a comparison of accuracy was conducted between the proposed method and the baseline method on the NTU-RGB+D 60 dataset for different actions. In the majority of actions, the proposed method outperformed the baseline method. Some actions exhibited significant improvements, as illustrated in Fig. 8, which displays the recognition accuracy improvements of the proposed method over the baseline. Notably, the "reading" action achieved a recognition accuracy improvement as high as 12.5%, while the "put on a shoe" action saw a 10% improvement compared to the baseline. Furthermore, actions such as "put on glasses," "headache," and "writing" also exhibited recognition accuracy improvements of 8.5%, 8%, and 7.9%, respectively.

As shown in Fig. 9, a selection of action screenshots is presented. In Fig. 9 (a) for the "reading" action, (b) for "headache," and (c) for "writing," these actions focus on localized hand movements and the connections between the hands and other joints. These connections encapsulate crucial spatial features that can be captured using the proposed connection feature extraction module. As expected, the actual results demonstrate significant improvements for these actions compared to the baseline. On the other hand, actions like (d) "putting on glasses," (e) "putting on shoes," and (f) "wearing a jacket" exhibit strong temporal dependencies. Leveraging the proposed long-distance frame difference feature extraction module effectively captures the long-term temporal

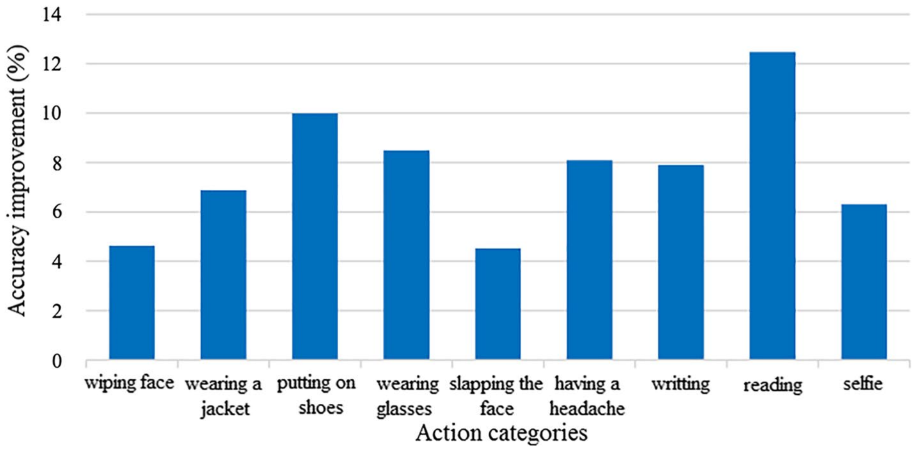
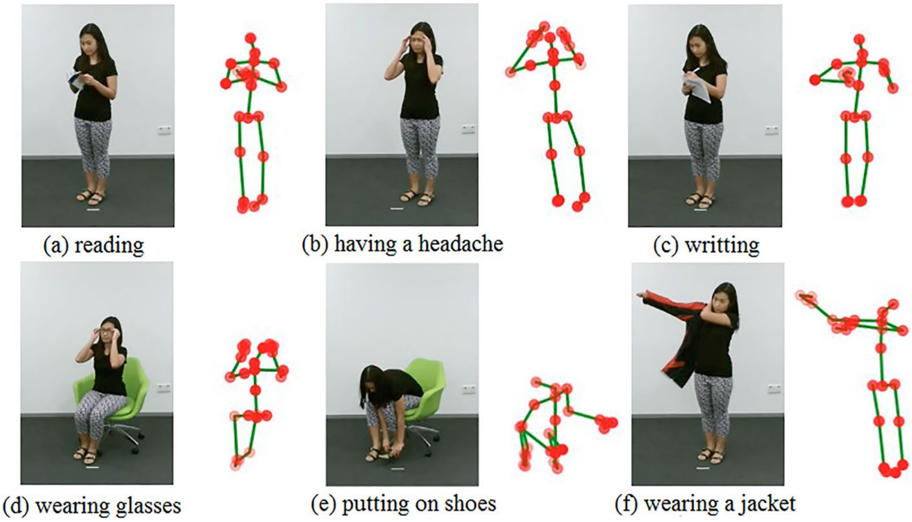


Fig. 8 Improved accuracy compared to the baseline on several actions in the NTU-RGB+D 60 dataset



(a) reading (b) having a headache (c) writting

(d) wearing glasses (e) putting on shoes (f) wearing a jacket

Fig. 9 Partial action demonstrations on NTU 60 dataset

Table 2 Dissolution experiments of multi-stream fusion

Method	Coordinate representation	Accuracy (%)
ST-GCN	Rectangular coordinates	88.93
Ours	Rectangular coordinates	89.57
Ours	Spherical coordinates	89.75
Ours	Rectangular + spherical coordinates	90.52

dependencies, as demonstrated by the results. These findings underscore the efficacy of the method proposed in this study.

Furthermore, to take it a step further, this paper introduces a dual-stream fusion network. Initially, the baseline method of this paper is subjected to dual-stream fusion, resulting in an accuracy of 88.93%, as indicated by ST-GCN in Table 2. This represents a notable improvement of 2.6% compared to the single-stream results, demonstrating a substantial enhancement. The proposed method conducts dual-stream fusion experiments separately in the two coordinate systems, with the results presented in Table 2.

As shown in Table 2, the fusion results of our method in the Cartesian coordinate system achieve an accuracy of 89.57%, and in the spherical coordinate system, the accuracy reaches 89.75%. Both of these results show a significant improvement compared to single-stream networks. Our method also verifies the fusion results using both coordinate representations simultaneously, referred to as multi-stream fusion, which ultimately achieves an accuracy of 90.52%.

4.4 Comparative experiments with existing methods

Firstly, this paper conducts experiments on the NTU-RGB+D 60 dataset to compare the proposed human behavior recognition method based on coordinate transformation and connection features with classic methods, including traditional approaches from the early days and recent advanced algorithms. Table 3 presents the comparative

Table 3 Comparative experiments conducted on the NTU-RGB+D 60 dataset

Method	Cross-subject (%)	Cross-view (%)
H-RNN [7]	59.1	64.0
ST-LSTM [29]	69.2	77.7
Two-Stream RNN [30]	71.3	79.5
STA-LSTM [8]	73.4	81.2
Ensemble TS-LSTM [31]	74.6	81.3
TCN [12]	74.3	83.1
GCA-LSTM [32]	76.1	84.0
Clips + CNN + MTLN [33]	79.6	84.8
VA-LSTM [34]	79.2	87.6
Ind-RNN [35]	81.8	88.0
ST-GCN (baseline) [14]	81.5	88.3
ARRN-LSTM [36]	80.7	88.8
HCN [13]	86.5	91.1
PR-GCN [37]	85.2	91.7
SR-TSL [38]	84.8	92.4
TS-SAN [39]	87.2	92.7
AR-GCN [40]	85.1	93.2
PB-GCN [41]	87.5	93.2
MANs [42]	82.7	93.2
AS-GCN [15]	86.8	94.2
2s-AGCN [16]	88.5	95.1
AGC-LSTM [43]	89.2	95.0
GCN-NAS [44]	89.4	95.7
DGNN [45]	89.9	**96.1**
SGN [25]	86.6	93.4
Ours	**90.5**	95.8

Bold is used to highlight key findings or best results that are particularly noteworthy or important within the context of the study. These bolded entries serve to draw the reader's attention to crucial data points, facilitating easier interpretation and understanding of the tables

experimental results on the NTU 60 dataset. Under the cross-subject evaluation, the accuracy of our algorithm is 90.5%, which is the best among all the compared methods. Under the Cross-view evaluation, the recognition accuracy is 95.8%, slightly lower than the highest-performing DGCN method by 0.3%. This is because the DGCN method employs not only skeletal and skeleton-stream data but also two additional motion streams, which are not utilized in our approach. The significant improvements of our method compared to the baseline methods highlight the effectiveness of our proposed approach.

Table 4 presents the comparative results on the NTU 120 dataset. Due to the dataset's recent release, this paper conducted comparisons with only a subset of methods. For the proposed algorithm, the recognition accuracy under the cross-subject evaluation reached 85.6%, and under the cross-setup evaluation, it achieved 87.4%. These accuracies surpass those of the baseline method proposed in this paper, as well as other state-of-the-art algorithms.

On the Kinetics dataset, the comparative experimental results of this paper are presented in Table 5. Due to the presence of significant noise in the Kinetics dataset and the reliance on the accuracy of pose estimation algorithms for obtaining skeleton

Table 4 Comparative experiments conducted on the NTU-RGB+D 120 dataset

Method	Cross-subject (%)	Cross-view (%)
Part-Aware LSTM [26]	25.5	26.3
ST-LSTM [29]	55.7	57.9
GCA-LSTM [32]	58.3	59.2
TSRJI [46]	67.9	62.8
SGN [25]	79.2	81.5
Poincare-GCN [47]	80.5	83.2
MV-IGNET [48]	83.9	85.6
FGCN [49]	85.4	87.4
Ours	**85.6**	**87.4**

Bold is used to highlight key findings or best results that are particularly noteworthy or important within the context of the study. These bolded entries serve to draw the reader's attention to crucial data points, facilitating easier interpretation and understanding of the tables

Table 5 Comparative experiments conducted on the Kinetics dataset

Method	Cross-subject (%)	Cross-view (%)
Feature Enc [50]	14.9	25.8
Deep LSTM [26]	16.4	35.3
Temporal Conv [12]	20.3	40.4
ST-GCN [14]	30.7	52.8
AR-GCN [40]	33.5	56.1
PR-GCN [37]	33.7	55.8
Ours	**33.8**	**57.0**

Bold is used to highlight key findings or best results that are particularly noteworthy or important within the context of the study. These bolded entries serve to draw the reader's attention to crucial data points, facilitating easier interpretation and understanding of the tables

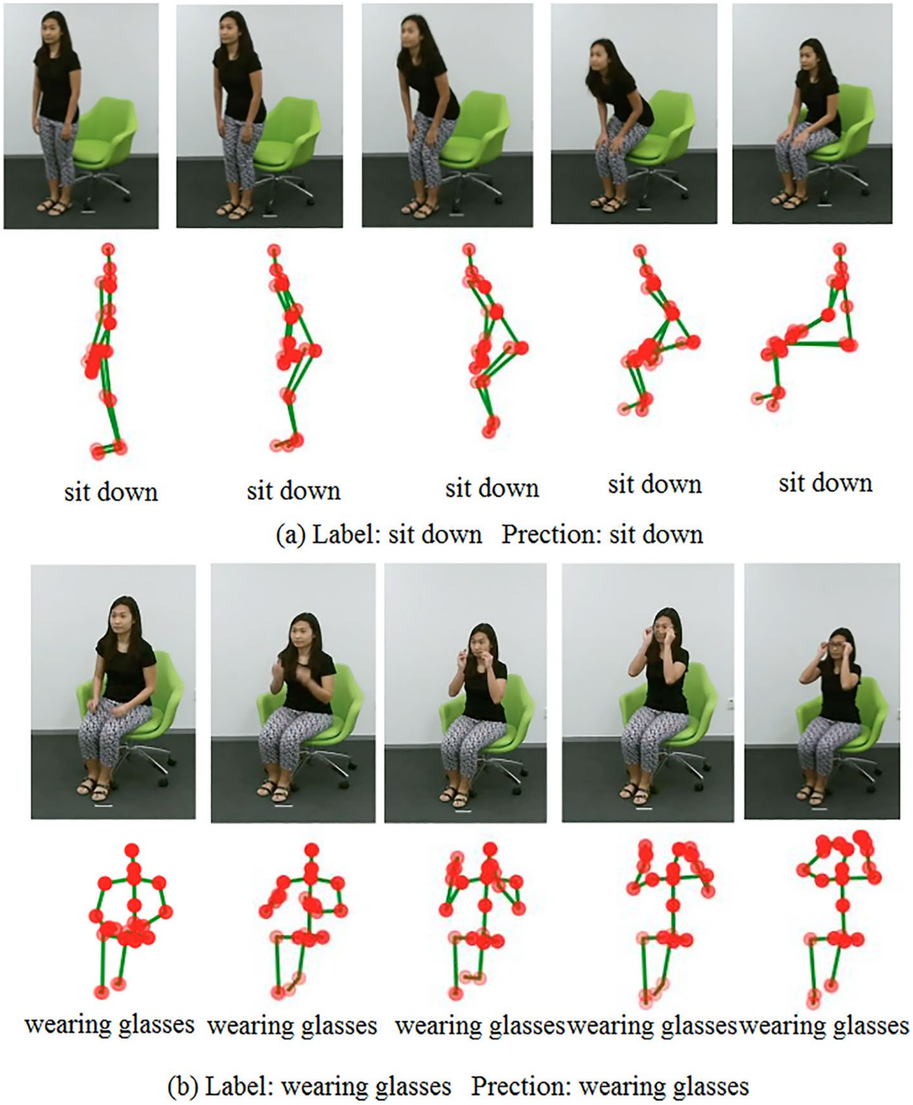
point data, the recognition accuracy on this dataset tends to be generally low, as reflected in the table. The proposed method achieved a top-1 accuracy of 33.8% and a top-5 accuracy of 57.0%.

Based on the results of the comparative experiments presented above, it can be observed that the human behavior recognition method proposed in this paper, based on coordinate transformation and connection features, has achieved advanced performance on multiple datasets. In comparison with many existing methods, the proposed approach demonstrates superior recognition effectiveness. This clearly emphasizes the validity and innovation of the work presented in this paper. The experimental results validate the effectiveness of the three proposed modules individually, affirming that these modules indeed perform as anticipated.

The NTU-RGB+D 60 dataset is widely recognized as a challenging benchmark for skeleton-based action recognition, with a diverse set of actions and cross-subject evaluation protocol that tests the generalization capabilities of models. Our method outperforms several state-of-the-art approaches on this dataset, as reported in Table 3, further highlighting the significance of the achieved improvements. While incremental improvements may seem modest in isolation, they can have a compounding effect when combined with other novel components, as demonstrated by our integrated framework. Beyond quantitative improvements, our work introduces novel methodological contributions, such as the attention-based modeling of implicit joint connections, long-range temporal feature extraction, and coordinate transformation, which advance the field of skeleton-based action recognition. The consistent performance improvements observed across multiple datasets (NTU-RGB+D 60, NTU-RGB+D 120, and Kinetics) further validate the robustness and generalizability of our approach.

4.5 Visualization results

As shown in Fig. 10, the recognition results of the proposed algorithm on the NTU dataset are illustrated. The upper portion of Fig. 10 displays several frames extracted from the captured videos, while the lower portion shows the corresponding skeleton



(a) Label: sit down Prection: sit down

(b) Label: wearing glasses Prection: wearing glasses

Fig. 10 Visualization results

point data alongside their predicted results. It is evident that the recognition outcomes of the proposed algorithm are accurate and aligned with the ground truth.

5 Discussion

This paper presents a novel approach to human behavior recognition through the integration of innovative techniques and modules. Its key contributions include:

(1) *Integrated framework* Our work introduces a comprehensive and innovative framework that seamlessly integrates three distinct techniques: coordinate transformation, connection feature extraction, and long-distance frame difference analysis. This integrated approach represents a significant advancement in the field of human behavior recognition, as it enables a holistic and multi-faceted representation of spatial and temporal features. By combining these complementary techniques, our framework achieves a synergistic effect, leading to improved recognition accuracy and robustness.

(2) *Coordinate transformation* A key novelty of our approach lies in the introduction of a coordinate transformation module. Traditionally, human skeletal data has been represented using Cartesian coordinates, which may fail to capture the nonlinear relationships and interconnections between joints effectively. Our work pioneers the use of spherical coordinates in the context of human behavior recognition, enabling a more accurate representation of the intricate joint relationships and body dynamics. By converting the Cartesian coordinates to spherical coordinates, our method can capture the nonlinear spatial characteristics of human movements, leading to enhanced feature representation and improved recognition performance.

(3) *Connection feature extraction* Another unique aspect of our work is the development of a connection feature extraction module. While previous methods have primarily focused on using joint connections as the graph structure for convolutional neural networks, our approach recognizes the inherent spatial feature information embedded within these connections. By explicitly extracting and leveraging these connection features, our method can capture the local spatial relationships between joints, providing valuable insights into the structural dynamics of the human body during motion. This novel approach complements the global spatial features extracted by graph convolutional networks, leading to a more comprehensive and accurate representation of human actions.

(4) *Long-distance frame difference analysis* Temporal dependencies play a crucial role in human behavior recognition, as actions unfold over time. While traditional methods have relied on short-range temporal convolutions or adjacent frame differences, our work introduces a long-distance frame difference feature extraction module. This novel technique enables the capture of long-term temporal dependencies by analyzing frame differences across extensive time intervals. By considering these long-range temporal patterns, our method can better understand the evolution of actions over time, leading to improved recognition accuracy, particularly for complex and prolonged movements.

(5) *Multi-stream fusion* To leverage the strengths of different coordinate representations, our work incorporates a novel multi-stream fusion approach. By fusing the

features extracted from both Cartesian and spherical coordinate systems, our method can capitalize on the complementary advantages of each representation. The multi-stream fusion technique enables our framework to capture a diverse range of spatial and temporal cues, leading to a more comprehensive and robust understanding of human actions. This innovative fusion approach represents a significant advancement in the field, as it combines multiple perspectives to achieve enhanced recognition accuracy.

In summary, this paper's contributions lie in its novel framework that seamlessly integrates coordinate transformation, connection feature extraction, and long-distance frame difference. These innovations collectively empower the method to achieve superior accuracy in human behavior recognition tasks, showcasing its potential impact on the field.

While this paper presents significant advancements in human behavior recognition, it also has certain limitations:

(1) *Dependency on data quality* The accuracy of the proposed method relies on the accuracy of the input data, particularly the quality of estimated joint positions. Noisy or imprecise joint data could affect the performance of the method.

(2) *Computational complexity* The introduced modules, especially the long-distance frame difference feature extraction, may increase the computational requirements of the method, potentially limiting its real-time applications on resource-constrained devices.

The paper opens up possibilities for further research and improvement in the field of human behavior recognition:

(1) *Robustness to noisy data* Future work could explore techniques to make the proposed method more robust to noisy input data, such as joint position estimation errors.

(2) *Efficiency improvements* Investigate methods to optimize the computational complexity of the proposed framework, making it more suitable for real-time applications.

(3) *Generalization* Extend the method's applicability to different domains, such as recognizing behaviors in animals or objects, by adapting and refining the proposed framework.

(4) *Transfer learning* Explore the potential of transfer learning to adapt the proposed method to new datasets or different sensor modalities, thus reducing the need for extensive data collection.

The ST-GCN method, proposed by Yan et al. [14] in 2018, has been widely recognized as a pioneering work in the field of skeleton-based action recognition. It introduced the concept of applying GCNs to skeletal data, representing the human body as a graph structure with joints as nodes and natural connections as edges. This approach effectively captures the spatial relationships between joints and enables the extraction of discriminative features for action recognition.

Despite its introduction several years ago, ST-GCN remains a highly influential and widely adopted baseline method in the field. Its impact can be attributed to several factors:

(1) *Simplicity and interpretability* ST-GCN's architecture is relatively straightforward, making it easy to understand and implement. This simplicity has facilitated its adoption by researchers and practitioners, fostering further exploration and development in the field.

(2) *Solid performance* Even with its simplicity, ST-GCN has demonstrated competitive performance on benchmark datasets, such as NTU-RGB+D and Kinetics. Its ability to achieve state-of-the-art results at the time of its introduction has contributed to its widespread recognition and adoption.

(3) *Extensibility* ST-GCN has served as a foundation for numerous subsequent works, with researchers proposing extensions and modifications to improve its performance or adapt it to specific scenarios. This extensibility has made ST-GCN a valuable starting point for further research in the field.

(4) *Reproducibility* The authors of ST-GCN have made their code and implementation details publicly available, enabling researchers to reproduce their results and build upon their work. This transparency has facilitated fair comparisons and accelerated progress in the field.

While several advanced methods have been proposed since the introduction of ST-GCN, it continues to be widely used as a baseline for comparison and evaluation in recent studies. Its simplicity, solid performance, extensibility, and reproducibility have solidified its position as a foundational work in the field of skeleton-based action recognition.

6 Conclusions

This paper presents a novel approach for human behavior recognition based on the combination of coordinate transformation and connection features. By extracting both local spatial features through the connection feature extraction module and long-term temporal dependencies using the long-distance frame difference feature extraction module, the proposed method effectively captures intricate patterns in human movements. The introduction of the coordinate transformation module further enhances the representation of distinct features. The integration of multi-stream fusion contributes to achieving superior recognition accuracy. Through extensive experimentation and comparison with existing methods on multiple datasets, the proposed approach consistently outperforms baseline methods and demonstrates its effectiveness in various scenarios. The visualizations of recognition results also provide tangible evidence of the method's success. In summary, this paper offers a comprehensive solution to the complex task of human behavior recognition, combining innovative techniques to achieve substantial improvements in accuracy.

Abbreviations

HOG	Histogram of oriented gradients
HoF	Histogram of flow
RNNs	Recurrent neural networks
LSTM	Long short-term memory

CNNs	Convolutional neural networks
GCNs	Graph convolutional networks
ST-GCN	Spatiotemporal graph convolutional network
CAE	Contrastive action encoding
GRU	Gated recurrent unit

Acknowledgements
Not applicable.

Author contributions
Conceptualization was done by WZ; methodology was done by JX; software was done by QW and YW; validation was done by RZ and JZ; writing was done by JX and JZ; writing—review and editing was done by FL; project administration was done by SD; funding acquisition was done by SD and WZ. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by the Natural Science Foundation of Fujian Province (Grant Nos. 2021J011086, 2023J01964, 2023J01965, 2023J01966), by the Fujian Province Chinese Academy of Sciences STS Program Supporting Project (Grant no. 2023T3084), by the External Collaboration Project of Science and Technology Department of Fujian Province (Grant No. 2023I0025), by the Guidance Project of the Science and Technology Department of Fujian Province (Grand No. 2023H0017), by the Xinluo District Industry-University-Research Science and Technology Joint Innovation Project (Grand Nos. 2022XLXYZ002, 2022XLXYZ004), and by the Qimai Science and Technology Innovation Project of Wuping Country.

Availability of data and materials
The datasets used and/or analyzed during in current study are available from the corresponding author on reasonable requests.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no conflict of interest.

Received: 1 December 2023 Accepted: 23 April 2024
Published online: 07 May 2024

References

1. A. Barkoky, N.M. Charkari, Complex Network-based features extraction in RGB-D human action recognition. J. Vis. Commun. Image Represent. **82**, 103371 (2022)
2. R. Yue, Z. Tian, S. Du, Action recognition based on RGB and skeleton data sets: a survey. Neurocomputing **512**, 287–306 (2022)
3. I. Laptev, On space-time interest points. Int. J. Comput. Vis. **64**, 107–123 (2005)
4. S. Dalal, V.P. Vishwakarma, S. Kumar, Feature-based sketch-photo matching for face recognition. Procedia Comput. Sci. **167**, 562–570 (2020)
5. O. Oreifej, Z. Liu, Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716–723 (2013)
6. R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 588–595 (2014)
7. Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1110–1118 (2015)
8. S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. IEEE Trans. Image Process. **27**(7), 3459–3471 (2018)
9. P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, N. Zheng, EleAtt-RNN: adding attentiveness to neurons in recurrent neural networks. IEEE Trans. Image Process. **29**, 1061–1073 (2019)
10. S. Qiu, T. Fan, J. Jiang, Z. Wang, Y. Wang, J. Xu, N. Jiang, A novel two-level interactive action recognition model based on inertial data fusion. Inf. Sci. **633**, 264–279 (2023)
11. C.Y. Ma, M.H. Chen, Z. Kira, G. AlRegib, TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition. Signal Process. Image Commun. **71**, 76–87 (2019)
12. T.S. Kim, A. Reiter, Interpretable 3d human action analysis with temporal convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–28 (2017)
13. C. Li, Q. Zhong, D. Xie, S. Pu, Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 786–792 (2018)

14. S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in *Thirty-second AAAI Conference on Artificial Intelligence*, vol 32, no 1, pp. 7444–7452 (2018)
15. M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3595–3603 (2019)
16. L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035 (2019)
17. K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 183–192 (2020)
18. Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 143–152 (2020)
19. N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, Z. Gong, Unsupervised representation learning with long-term dynamics for skeleton based action recognition, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol 32, no 1, pp. 2644–2651 (2018)
20. L. Lin, S. Song, W. Yang, J. Liu, Ms2l: multi-task self-supervised learning for skeleton based action recognition, in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2490–2498 (2020)
21. H. Rao, S. Xu, X. Hu, J. Cheng, B. Hu, Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. Inf. Sci. **569**, 90–109 (2021)
22. K. Su, X. Liu, E. Shlizerman, Predict and cluster: unsupervised skeleton based action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9631–9640 (2020)
23. Q. Nie, Z. Liu, Y. Liu, Unsupervised 3d human pose representation with viewpoint and pose disentanglement, in *European Conference on Computer Vision*, pp. 102–118 (2020)
24. L. Li, M. Wang, B. Ni, H. Wang, J. Yang, W. Zhang, 3d human action representation learning via cross-view consistency pursuit, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4741–4750 (2021)
25. P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1112–1121 (2020)
26. A. Shahroudy, J. Liu, T.T. Ng, G. Wang, Ntu rgb+ d: a large scale dataset for 3d human activity analysis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1010–1019 (2016)
27. J. Liu, A. Shahroudy, M. Perez, G. Wang, L.Y. Duan, A.C. Kot, Ntu rgb+ d 120: a large-scale benchmark for 3d human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2684–2701 (2019)
28. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, A. Zisserman, The kinetics human action video dataset (2017). arXiv:1705.06950
29. J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, in *Computer Vision-ECCV 2016: 14th European Conference*, pp. 816–833 (2016)
30. R. Zhao, H. Ali, P. Van der Smagt, Two-stream RNN/CNN for action recognition in 3D videos, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4260–4267 (2017)
31. I. Lee, D. Kim, S. Kang, S. Lee, Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1012–1020 (2017)
32. J. Liu, G. Wang, P. Hu, L.Y. Duan, A.C. Kot, Global context-aware attention lstm networks for 3d action recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1647–1656 (2017)
33. Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3d action recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3288–3297 (2017)
34. P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View adaptive recurrent neural networks for high performance human action recognition from skeleton data, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2117–2126 (2017)
35. S. Li, W. Li, C. Cook, C. Zhu, Y. Gao, Independently recurrent neural network (indrnn): building a longer and deeper rnn, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5457–5466 (2018)
36. W. Zheng, L. Li, Z. Zhang, Y. Huang, L. Wang, Relational network for skeleton-based action recognition, in *2019 IEEE International Conference on Multimedia and Expo*, pp. 826–831 (2019)
37. S. Li, J. Yi, Y.A. Farha, J. Gall, Pose refinement graph convolutional network for skeleton-based action recognition. IEEE Robot. Autom. Lett. **6**(2), 1028–1035 (2021)
38. C. Si, Y. Jing, W. Wang, L. Wang, T. Tan, Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network. Pattern Recognit. **107**, 107511 (2020)
39. S. Cho, M. Maqbool, F. Liu, H. Foroosh, Self-attention network for skeleton-based human action recognition, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 635–644 (2020)
40. X. Ding, K. Yang, W. Chen, An attention-enhanced recurrent graph convolutional network for skeleton-based action recognition, in *Proceedings of the 2019 2nd International Conference on Signal Processing and Machine Learning*, pp. 79–84 (2019)
41. K. Thakkar, P.J. Narayanan, Part-based graph convolutional network for action recognition (2018). arXiv:1809.04983
42. C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, J. Chen, Memory attention networks for skeleton-based action recognition. IEEE Trans. Neural Netw. Learn. Syst. **33**(9), 4800–4814 (2021)
43. C. Si, W. Chen, W. Wang, L. Wang, T. Tan, An attention enhanced graph convolutional lstm network for skeleton-based action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1227–1236 (2019)
44. W. Peng, X. Hong, H. Chen, G. Zhao, Learning graph convolutional network for skeleton-based human action recognition by neural searching, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, pp. 2669–2676 (2020)
45. L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with directed graph neural networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7912–7921 (2019)

46. C. Caetano, F. Brémond, W.R. Schwartz, Skeleton image representation for 3d action recognition based on tree structure and reference joints, in *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 16–23 (2019)
47. W. Peng, J. Shi, Z. Xia, et al. Mix dimension in poincarégeometry for 3d skeleton-based action recognition, in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1432-1440 (2020)
48. M. Wang, B. Ni, X. Yang, Learning multi-view interactional skeleton graph for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**(6), 6940–6954 (2023)
49. H. Yang, D. Yan, L. Zhang, Y. Sun, D. Li, S.J. Maybank, Feedback graph convolutional network for skeleton-based action recognition. IEEE Trans. Image Process. **31**, 164–175 (2021)
50. B. Fernando, E. Gavves, J.M. Oramas, A. Ghodrati, T. Tuytelaars, Modeling video evolution for action recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5378–5387 (2015)

Publisher's Note