Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **7**, Number **4**, November **2022**

A New Bone Direction Prediction Method Based on Spatial-temporal Graph Convolutional Network

Ya-Pei Feng

College of Computer Science and Technology
Hangzhou Dianzi University
No. 1, No. 2 Street, Baiyang Street, Qiantang District, Hangzhou,310018, China
fionayp717@163.com

Zhe-Ming Lu

School of Aeronautics and Astronautics
Zhejiang University
No. 38, Zheda Road, Xihu District, Hangzhou,310027, China.
zheminglu@zju.edu.cn

*Corresponding author: Zhe-Ming Lu

Received March 11, 2022, revised April 22, 2022, accepted July 17, 2022.

Abstract. *Human pose estimation is an important research area in computer vision, and is a pre-requisite task for research on human motion recognition, behavior analysis, and human-computer interaction. A dynamic skeleton model can be naturally represented as a series of time series of human joint positions. To express the real-time motion state of skeletal nodes during human motion, graph convolutional neural networks are the most suitable choice because the human skeleton is most similar to the graph. In this article, given the neglect of skeletal motion direction by previous methods, we propose a new bone motion recognition method that considers bone direction calculation and prediction. Based on the motion direction of the bone,the skeletal orientation is defined as the subtraction of the coordinates of the relevant skeletal points. Firstly, the bone points are obtained through the attention mechanism, and we take two points with a connection relationship are taken to express the predicted information between two joints of the bone. The prediction yields nodes as new nodes, thus updating the graph network structure. Finally, all the generated new nodes form a new skeleton, which improves the information and discriminative power of the skeletal representation.*

Therefore, our algorithm can not only express the coordinates of the joints, but also the prediction information between the two joints of the bone, which is more informative and discriminative. We selected two classical 3D skeletal motion recognition datasets:NTU-RBD and Kinetics datasets as our training and test subsets, and the comparison experiments show that our proposed algorithm performs better than other methods and improves the accuracy and efficiency of pose recognition.

Keywords: Action recognition, Bone direction prediction, Attention mechanism,Skeleton reconfiguration.

1. **Introduction.** The problem of pose estimation and action recognition is a popular research area in computer vision. This problem is a key to successful analysis of peoples' behaviors on video. Skeleton data based motion recognition and pose estimation are widely concerned and researched by worldwide researchers, not only because of their worldwide applications in video understanding and detection, but also because they are still challenging tasks in complicated backgrounds.

The dynamic skeletal modality can naturally represent the time sequence of human joint positions in two-dimensional space or three-dimensional space coordinates, and then its movement can be identified by dissecting the pattern of its motion posture [**?**, **?**].Action recognition applied to the skeleton is performed by adopting the joint coordinates at each time step to constitute a feature characteristics, and then perform time interpretation on them.The conventional RGB videos and deep learning-based methods artificially construct the skeleton as a spatial sequence oriented to the main body joints, which cannot accurately indicate the dependency between associated joints. Large scale 3D skeleton datasets are available for researchers to explore skeleton-based human action recognition and detection with the constantly evolving of low-cost human skeleton information acquisition systems [**?**]. However, the performance of these manual function-based methods is almost unsatisfactory because it cannot take into account all the influencing factors at the same time. Many works [**?**, **?**] explore the temporal evolution of actions based on long short-term memory (LSTM) based framework, and many improvements have been achieved in the literatures. The traditional deep learning-based method is to artificially abstract the skeleton as a series of joint coordinate vectors [**?**, **?**, **?**] or pseudo-images [**?**, **?**, **?**], and feed it into RNN or CNN to generate predictions.

In the era of big data, as deep learning techniques and artificial intelligence technologies become more and more mature, while the application scenarios we face become more and more complex, data-driven approaches gradually become the mainstream of pose recognition. Thus, the problem of human skeleton-based motion recognition focuses on two aspects: the first is the variability of the input data, such as scaling, rotation and translation, which can largely increase the completeness of the training data; the second is the modeling of human behavior, which enables a concrete description of the human pose. These two aspects are mutable, dynamic, and have similarities with each other. Most of the currently available skeleton-based action recognition approaches utilize a representation of the corresponding position of the joint coordinates, e.g., spatiotemporal pyramidal models and hidden Markov models.

Aiming at the demand of dynamic recognition, the most important point in building a structural model for action classification is to extract both visual characteristics of the appearance and temporal evolution of the joints. Researchers propose the representation of skeleton data with CNN [**?**], they applied the faster R-CNN based object recognition framework for action detection in the time domain to effectively extract epistemic and action features in regions of interest. In addition, in the work of paper [**?**], RNN models are also applied to provide abundant spatial domain characteristics for LSTM models by exploring the geometric relationships between joints and using different geometric feature streams for learning training. Many improvements have been made by scholars in order to more adequately represent the dependencies between the relevant joints. For the action recognition with skeleton task, in the work of [**?**], the GCN framework was chosen to represent the structure of the skeleton as a graph in non-Euclidean space by modeling the skeleton data, Where key skeletal points in the human body with connectivity are characterized as vertices of the graph, and temporal edges between corresponding joints in consecutive frames are characterized as edges of the graph.

The sampling function based on distance is adopted to construct the graph convolutional layer, and it is also applied as the fundamental component for constructing the complete ST-GCN. However, the shortcomings of the graph construction process in ST-GCN are: (1) The structure of GCN is divided into different layers, where each layer contains different multi-level semantic information. However, the ST-GCN model does not allow for complete and accurate modeling of the high-level semantic information contained in all levels. (2) The changes in skeletal orientation before and after limb movements are

not updated in time for samples with different movement categories. The movement directions of two points in the bones that have a connection relationship are taken into consideration [**?**]. Therefore, a fixed graph structure may not be optimal.

For the purpose of maintaining the data of the detection model updated with changes in skeletal orientation before and after limb movements, we will propose a network with temporal dimensions (Temporal) and spatial dimensions (Spatial): the spatial dimension is reflected in the skeleton diagram within a frame, and the time dimension is constructed to connect the nodes in the same position of the diagrams in adjacent frames, and then according to the skeleton, the sequence constructs a spatial-temporal graph. Then, a bone direction is artificially set, and the two points that have a connection relationship are taken as the new node according to the direction of the bone. Finally, all the generated new nodes form a new skeleton.

2. **Skeleton Graph Construction.** The structure of the graph is generally very irregular and can be considered as a kind of infinite-dimensional data, so it has no translation invariance,so the surrounding structure of each node may be unique,which makes traditional CNN and RNN invalid instantly. In order to deal with this type of data, many related research works have emerged, such as Graph Neural Network, DeepWalk, node2vec,and so on. GCN can extract features from the graph data, and then we can perform node classification, graph partitioning and classification, and link relationship prediction by the features of graph data.

The recognition of human skeletal structure is based on action recognition in two dimensions of time and space. For pictures, when doing target recognition, what we care about is actually only a part of the key points on the two-dimensional picture. The video is composed of frame by frame images, and the human skeleton data of each frame of the video is made up of the feature data of different human joints. Therefore, it is necessary to apply for a convolutional neural network based on graph structure to characterize the dynamic skeleton data. As we all know, deep learning-based research has gradually generalized to neural network structures for processing graph data due to the substantial increase in the amount of data and the emergence of complex scenarios. The authors propose a model based on a two-stream adaptive graph convolutional network (2s-AGCN) [**?**], in which the topology of the graph can be learned uniformly or individually in an end-to-end manner by BP algorithm implementation. Also, a dual-stream framework for modeling first level of visual information and second level of motion flow information is proposed in the paper, which has a significant improvement on recognition accuracy. The paper [**?**] proposes a new GCN trained based on reinforcement learning to solve the problem of action recognition by using joint relational inference, which has practical applications.

In this paper, we propose to design a representation of a skeleton sequence applicable to human action recognition by constructing a graph neural network through time and space modeling as a spatio-temporal graph model based on the ST-GCN model. Also, since our model incorporates the prediction of skeletal orientation, we define this model as the spatio-temporal graph model with orientation prediction (DPST-GCN).

As shown in Figure 1, the ST-GCN model is formulated on a sequence of skeleton diagrams, where every single node corresponds to a joint in the human body. The joints in a frame are connected to the edges according to the connectivity of the human body structure. And the video is decomposed into frame by frame, and in each frame , a spatial graph is constructed based on the distribution characteristics of the real skeletal interconnections of the human body; meanwhile, the key points in the same location and function in two adjacent frames are connected to form a spatial-temporal edge. Thus, the two categories of edges in the model are: spatial edges matching the raw connections of

joints and temporal edges connecting the same joints in successive time strides. Eventually, different frames can be convolved along the spatial and temporal dimensions to jointly build a multilayer spatio-temporal graph convolution model. The two dimensions of this spatio-temporal graph convolution model can also be understood as: (1) Temporal dimension: the nodes at the same position of adjacent frames in the graph, and the points at the same position of the two preceding and following frames as the points to be involved in the convolution. (2) Spatial dimension: the skeleton diagram in one frame, a convolution center point is determined first, and the points adjacent to it are taken as the points to be involved in the convolution.

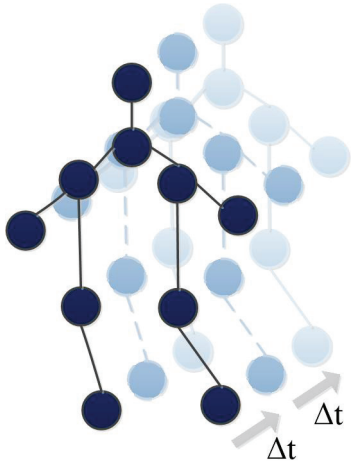

FIGURE 1. The spatial temporal graph of a skeleton sequence

In addition, in our DPST-GCN model, we added the attention branch to put more emphasis on categorical and discriminative features, while keeping the common features about joint electricity in another branch to maintain the integrity of the features.

3. Proposed Scheme.

3.1. Spatial and temporal graph convolutional networks(ST-GCN).

ST-GCN [**?**] was proposed to adapt to the dynamic environment and complex background of the pose recognition process, and to be able to present the hierarchical representation of the skeleton sequence as a spatial-temporal map. In order to describe the skeleton sequence with N nodes and T frames in a concrete form, the algorithm constructs an undirected spatial and temporal graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$, and it is also necessary to characterize each joint point of the human body and the connections between adjacent frames. In the graph, the set of all joint nodes including the entire skeleton sequence is represented by $\mathbf{V} = \{v_{ti} | t = 1, 2, ...T, i = 1, 2, ...N\}$, where T is the number of features of each node. Each node in the graph has its unique feature vector, and each edge in the set $\mathbf{E}$ connects the adjacent joints of the human body in each frame, while ensuring that each edge connects the same joints between consecutive frames.

Since the human body accomplish a entire action during motion is combined with frames of different lengths in the video, the information of skeletal nodes in each frame needs to be characterized by spatial data, while the information between different frames needs to be represented by a temporal graph. Therefore, we need to utilize a spatial-temporal graph to model the structured information between these joints along the spatial and temporal dimensions.

The ST-GCN algorithm adopts the idea of convolutional network. In the skeleton node network, a convolution center is first determined, and then in the spatial dimension, the adjacent point is taken as the point that needs to participate in the convolution; at the same time, in the temporal dimension, take the points where the front and back frames are at the same position as the points that need to participate in the convolution. Finally, after determining the convolution center and the points that need to participate in the convolution center, the convolution operation is performed, which can aggregate the information of the different temporal dimension and the complex spatial dimension at the same time.

In our proposed scheme, the node set defined as $\mathbf{V} = \{v_{ti}|t = 1, 2, ...T, i = 1, 2, ...N\}$ represents the whole joints in a skeleton sequence. The body's physical bone connection ES and the different time joint trajectories $E_F = \{v_{ti} \times v_{(t+1)i}\}$ constitute the edge set $\mathbf{E}$ of the space-time graph. The feature vector $F(v_{ti})$ on each node is a three-dimensional space coordinate vector. When the skeleton sequence is the result of video estimation of human body pose, the feature vector of the node is the two-dimensional space coordinate vector and the confidence of the node is estimated.

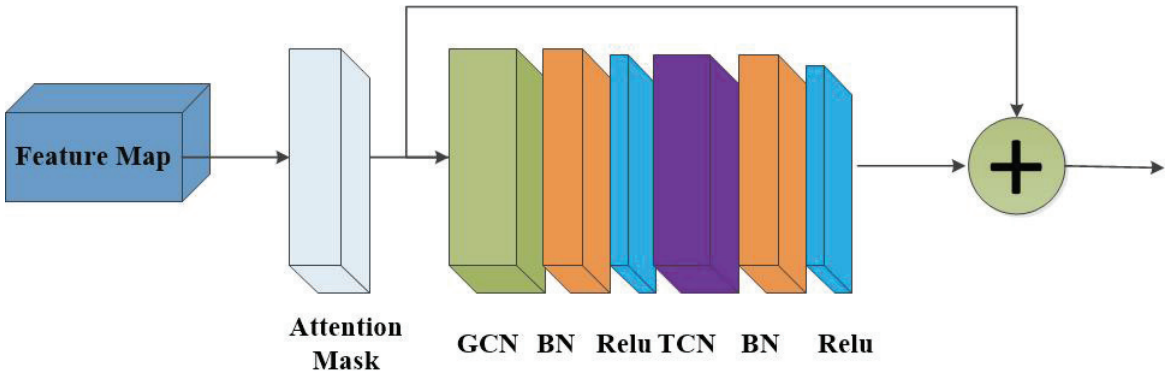

FIGURE 2. The structure of the spatial-temporal block

For the individual node in the graph, it can be characterized by the corresponding feature vector, for example, for a node i in a frame T of video clip, its graph convolution can be described by the formula defined in Eq. (1).

$$f_{out}(v_i) = \sum_{v_j \in B_{v_{ti}}} = f_{in}(v_{ij}) * w(v_{ti}) \tag{1}$$

where $f_{in}(v_{ij})$ is the feature vector of the node set in the graph v_{ij} in the input feature map, and $B_{v_{ij}}$ is a collection of the nodes within a certain distance away from the node v_{ti} in the space and time domains, which can be written as:

$$B(v_{ti}) = \{v_{ti}|d(v_{tj}, v_{ti}) \leq D\} \cup \{v_{qi}| \, |q - t| \leq \tau/2\} \tag{2}$$

And w can be thought of as a weighting constructor in the convolution operation that performs the weighting operation, which can offer a suitable weight vector based on a specific input. In order to make the disordered and unfixed number of neighboring nodes become orderly, the ST-GCN algorithm adopts the method of graph label processing, and divides the neighborhood into a certain number of M subsets according to the segmentation strategy, and the labels of the nodes in each subset are the same, so $B(v_{ti})$ can be mapped to labels $L = \{0, 1, 2...M - 1\}$ .Finally, the weighting function of graph

convolution operation can be defined as:

$$w(v_{ti}, v_j) = w(l(v_{tj})) \tag{3}$$

Among them, $l(\cdot)$ is the label mapping function determined by the segmentation strategy. From the temporal dimension, when the label mapping $l(v_{ti})$ of a node is confirmed, the label of the node at other times is also confirmed:

$$l(v_{qi}) = l(v_{ti}) + (q + \tau/2) \times K \tag{4}$$

where K is the number of labels of neighborhood nodes v_{ti} in a single time.

As in the structure of the spatial-temporal block shown in Figure 2, the feature maps are first fed to the full attention module to mater the corresponding attention masks. The learned attention mask represents the feature weights of the corresponding feature maps, so that the mask has the same size as the corresponding feature maps of the input data. It aims to assign the much higher weight values to more discriminative features. On the one hand, the attention mask is multiplied by the corresponding feature maps one by one, and the obtained results are added to the feature maps one by one as the input feature maps for the next layer of image convolution.On the other hand, the result of the element-by-element multiplication of the attention mask and the input feature map is transferred to the attention branch to participate in the loss calculation of the attention branch. Thus, the attention mechanism can be applied to the attention branch of the graphical convolutional network branch as well as the attention branch of the multitasking framework.

3.2. **Bone direction graph construction.** In this section, we will introduce our proposed DPST-GCN model. The input data of skeletal sequence action recognition continuously records the characteristic information of the main joint points of the human body. The 3D spatial location is applied as the feature information, and each joint point of the human skeleton in each frame of the video corresponds to a 3D feature vector. The skeleton diagram of the node set is shown in Figure 3.To more clearly describe the body's joint points, we used a configuration of 25 body joints from the 3D dataset, with the coordinates and locations of the different osteoarticular nodes characterized by numbers. In a certain order [**?**] these joints are labeled as : 1-bottom of spine, 2-middle of spine, 3-neck, 4-head, 5-left shoulder, 6-left elbow, 7-left wrist, 8-left hand, 9-right shoulder, 10-right elbow, 11-right wrist, 12-right hand, 13-left hip. 14-left knee, 15-left ankle, 16-left foot, 17-right hip, 18-right knee, 19-right ankle, 20-right foot , 21-Spine, 22-Left hand tip, 23-Left thumb, 24-Right thumb and 25-Right thumb.

Then we divide the skeleton graph of the node set into different subsets according to the division of the space configuration to divide the neighbors.

In ST-GCN [**?**], all the keypoints extracted from each frame form a key node set, and the same key points in two adjacent frames are connected to form all edges of the timing edge. This constitutes an edge set (edge set), that is, a required space-time diagram, which naturally retains the spatial information of the key points of the skeleton, and enables the trajectory of the key points to be expressed in the form of time series edges.

However, ST-GCN is able to express the feature flow of the skeleton data well by building feature vectors on each vertex in the graph structure. However, these visual feature streams contain only 2D or 3D coordinates of the joints, and do not characterize the skeletal features of the motion relationship between two joints. In the actual experimental process, we found that for the recognition of dynamic skeletal features, the length and motion direction of the skeleton are more informative and discriminative.

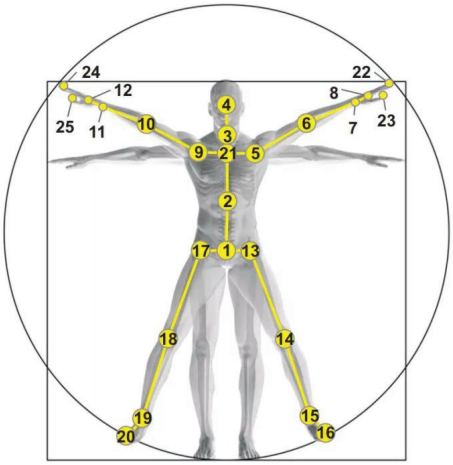

FIGURE 3. The joint labels of the NTU-RGBD dataset

Hence, in order to explore the directional motion information of the skeleton data as a way to represent the skeletal directional feature flow, we need to predict the motion direction of each skeletal point in each frame by the distance and relationship information between adjacent skeletal nodes, so as to continuously update the directional state of the skeleton. Just as shown in Figure 4, the blue mark represents the constantly updated bone direction. The specific steps are as follows:

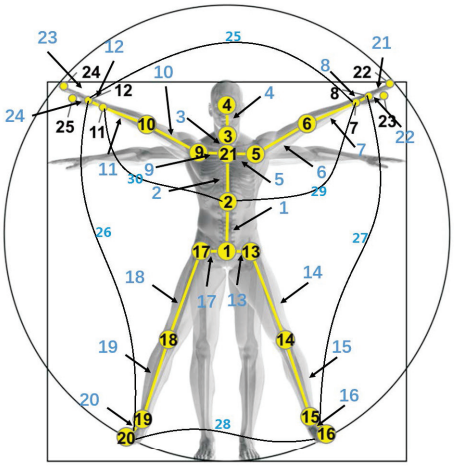

FIGURE 4. The prediction of the bone direction.The node serial number marked in black is the initial label, and the new nodes generated by one iteration are marked in blue

Firstly, given an input sample, we statistically obtain the skeletal data based on the distribution of the joint points.

Secondly, since different parts of the body interact with each other when humans move, meaning that the joints of the body parts interact and move together. Prediction of more complex motion trajectories can be achieved based on the combination of 3D trajectories of node movements. We artificially set a skeleton direction, take the two points with a connection relationship as the inner and outer points, and take the middle position of the

two nodes as the new nodes according to the skeleton direction; then all the generated new nodes constitute the new skeleton. This processes can be described as follows:

(1) The number of main nodes of the skeleton is 24,which can be described as $num\_node = 24$,and the connection between two nodes can be represented as $self\_link = (i, i)$ [**?**]. Here we show the the original stgcn joint method, original bone-based method and different direction bone-based method, which are represented as the edge format: (origin, neighbor) in the following Table 1.

(2) The neighbour node can be described as $neighbor = inward + outward$, and the inward and the outward can be described as $inward = (i - 1, j - 1)$ for (i, j) in inward_orientation_index, the original outward node is named as $outward = (j, i)$ for (i, j) in inward, we change the outward node as $outward = (i - 1, j - 1)$ for (i, j) in outward_ori_index, just as the example shown in Table 2. Then The coordinates of two adjacent skeletal points are subtracted from each other to give the orientation of the new skeleton.

For example, given a bone with its source joint $v_1 = (x_1, y_1, z_1)$ and its target joint $v_2 = (x_2, y_2, z_2)$, the vector of the bone is calculated as $e_{v1,v2} = (x_2 - x_1, y_2 - y_1, z_2 - z_1)$.

(3) Since DPST-GCN requires sharing weights at distinct nodes, it is essential to keep the scale of input data consistent across nodes. Through multiple iterative operations,the generation of skeletal frames is achieved by point-to-edge and edge-to-point conversions.

Then, all the direction vectors are fed to the GCN, and the direction of the bones is predicted by the directional difference of adjacent skeletal points, which enables the prediction of action labels.

In Table 1 we show the link based, original skeletal point information used in the stgcn network, and at the same time, information about the skeletal points calculated using the node directions described in this paper, e.g., skeletal direction 1 to skeletal direction 4. As can be seen from the table, the predicted directions based on the skeletal motion directions are chosen to express the detailed information about the human skeletal points more completely. Since the projection of skeletal directions is done based on the inner and outer nodes and the link relationship between the nodes. Table 2 shows the information of different inward and outward orientation indices .

Finally, similar to the bone stream information, the coordinates of the bone points are subtracted as the direction of the bone, and the direction vector is sent to the adaptive GCN to predict the action label. Furthermore, we incorporate with the attention graph model with the pooling operation [**?**] to retain the local properties of different bones and the graph structures . Meanwhile, we perform the graph convolution operations on the feature vector of the bone point and the feature vector of the motion direction to further improve the performance. Then the bone stream and the bone direction stream are sent to the network together, just as the structure shown in Figure.5.

4. **Experiments.**

4.1. **Datasets.**

To perform the comparison with the state-of-are, we select two classical action recognition datasets as experimental data: the skeleton sequence dataset (NTU-RGB+D) [**?**] and the Kinetics skeleton [**?**],which are both the large-scale datasets.

(1) NTU-RGB+D: This data set records the three-dimensional coordinate positions of 25 joint points of the human body in continuous time (X, Y, Z). It contains 56880 action samples and a total of 60 actions types, including 50 single-player action types and 10 two-player action types. The data set is classified into two subsets: cross-object (X-Sub) and cross-perspective (X-View). In the X-Sub, the training set and the test set are executed by different people. The training set has 20 action performers and a total of

Table 1. Different directions of the bone in edge format.

The original stgcn joint method	[(1, 2), (2, 21), (3, 21), (4, 3), (5, 21), (6, 5), (7, 6),(8, 7), (9, 21), (10, 9),(11, 10), (12, 11), (13, 1),(14, 13), (15, 14), (16, 15), (17, 1), (18, 17), (19, 18),(20, 19),(22, 23), (23, 8), (24, 25), (25, 12)]
The original bone-based method	[(1 , 2),(2 , 3), (3 , 4), (5 , 6), (6 ,7), (7 , 8), (5, 9),(9 ,10), (10, 11), (11, 12), (1 , 13),(13, 14), (14, 15),(15, 16),(1 ,17), (17, 18),(18, 19),(19, 20), (8 ,21),(8,22), (21, 22), (12, 23), (12, 24), (23, 24), (13, 17), (2 , 5), (2 , 9), (3 , 5), (3 , 9)]
The direction1 bone-based method	[(23,24), (23, 12), (24,12), (12, 11), (11 ,10), (10 , 9),(9, 5),(9 , 3), (9, 2), (3, 4), (5 , 6), (6, 7), (7, 8),(8 , 21),(8 , 22), (21, 22), (2, 1), (1, 17),(17,18),(18, 19), (19,20), (1, 13), (13 , 14), (14 , 15), (15 , 16)]
The direction2 bone-based method	[(9 ,10), (10,11), (11,12), (12,24), (12,23), (5 , 6), (6 , 7), (7 , 8), (8 ,21), (8 ,22), (17 , 18),(18 , 19), (19, 20),(13 , 14),(14,15), (15, 16),(3 , 4),(9 , 3),(9 , 2), (5, 3), (5,2),(13, 1), (17 , 1),(1 , 2)]
The direction3 bone-based method	[(10, 9), (11, 9), (12, 9), (23, 9), (24, 9), (6 , 5), (7 , 5), (8 , 5), (21, 5), (22, 5), (4 , 3), (18, 17), (19, 17), (20, 17), (14, 13), (15, 13), (16, 13)]
The direction4 bone-based method	[(2 , 9), (9 , 10), (10, 11), (11, 12), (12, 24), (12 ,23),(2 , 5), (5 , 6), (6 , 7), (7 , 8), (8 , 21), (8 , 22),(1 , 13), (17, 18), (18, 19), (19, 20),(1 , 17), (13, 14), (14, 15), (15, 16), (2 , 3), (3 , 4)]

Table 2. The inward orientation index in edge format.

Inward_orientation_index 1	[(1, 2), (2, 3), (3, 4), (6, 7), (7, 8), (5, 9), (10, 11), (11, 12), (1, 13), (14, 15), (15, 16), (1, 17), (18, 19), (19, 20), (8, 21), (8, 22), (12, 23), (12, 24), (23, 24), (21, 22), (13, 17), (2, 5), (2, 9), (3, 5), (3, 9), (5,6),(9,10), (13,14),(17,18), (5,8), (6,8), (9,12), (10,12),(13,16),(14,16), (17,20),(18,20)]
Inward_orientation_index 2	[(1, 2), (2, 3), (3, 4), (6, 7), (7, 8), (10, 11), (11, 12), (1, 13), (14, 15), (15, 16), (1, 17), (18, 19), (19, 20), (8, 21), (8, 22), (12, 23), (12, 24), (23, 24), (21, 22), (2, 5), (2, 9), (5,6), (9,10), (13,14),(17,18)]

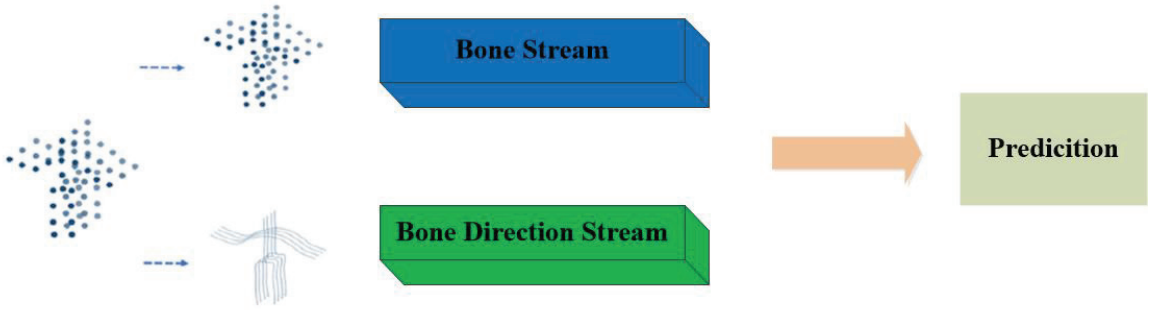

FIGURE 5. Illustration of the structure of the bone direction prediction

40,320 sample sequences, and the test set has 20 action performers and a total of 16,560 samples; While in X-View, the training set is the bone sequence of all people under the view of camera 2 and 3, and the test set is the bone sequence of all people under the view of camera 1. The training subset and test subset contains 37920 and 18960 samples, respectively.

(2) Kinetics: consisting of 400 actions with between 400 and 1150 video clips per action, with each video clip being approximately 10 seconds long. And the entire dataset is also partitioned into two subsets: the training subset (containing 240,000 clips) and the validation subset (containing 20,000 clips).

4.2. **Implementation Details.** We conduct the experiments on PyTorch deep learning framework with an AI computing platform with Nvidia A100 GPU and Aurora ParaStor300S parallel storage system. During the training process, we select the Adam optimizer with faster convergence rate, which is used to calculate the adaptive learning rate of each parameter with an initial value of 0.1 and a value of 0.0001 for the weight decay coefficient. Then, after the 10th epoch and the 50th epoch time, the learning rate is reduced to 1/10. In the back propagation structure of the neural network, in terms of the error back propagation and parameter iterative optimization perspective, the cross entropy is chosen as the loss function of the gradient. Brief description of the experimental steps is as follows.

Step 1: Data pre-processing. For videos containing human standing, walking, sitting, running, jumping, squatting, kicking, punching, waving and other actions, the whole multi-objective action monitoring system is built mainly relying on openpose's pose recognition environment. Different gesture features are detected and integrated as recognition features of the complete action through the openpose toolbox. To ensure that the skeleton sequence is fixed at 100 frames, we will zero fill if the length is not long enough, and conversely, if the length is larger, we will extract frames larger than 100 at equal intervals.

At the same time, for the NTU-RGBD data set, when utilizing the NTU-RGBD data set samples, we first perform frame padding, repetitive sampling, and normalization, and centering operations on each sample to achieve completeness and uniformity of the data samples.

Step 2: Training. The pre-processed data were input into our DPST-GCN model according to the skeletal orientation map generation method in Section 3.2, while the parameters were fine-tuned to ensure the stability and convergence of the model.

Step 3: Test. To ensure that the space represented by the test and training sets is the same, we preprocess the video to be predicted in the same way as in step 1. The preprocessed data is then fed into the trained network for classification and computation.

4.3. **Comparison with the State of the Art.** To verify the effectiveness of our algorithm, we select two classical datasets: the NTU-RGBD dataset [**?**] and the Kinetics-Skeleton dataset [**?**], and on the basis of these two datasets, we compare our model with several state-of-the-art methods associated with the bone point based action recognition for experiments. For the Kinetics-Skeleton dataset, we recorded the recognition accuracy for the top 1 and top 5. The action recognition comparative experimental outcome is given in Table 3 with RGB method [**?**], Feature map Enc. [**?**], Deep LSTM [**?**], Temporal Conv [**?**]and the ST-GCN [**?**] based alogrithm. And these algorithms are very representative experiments. Similarly, we also compared representative algorithms: spatial-temporal LSTM with trust gate (ST-LSTM+TS) [**?**], Clips CNN+multitask learning (C-CNN+MTLN) [**?**], and ST-GCN [**?**]. It should be noted that the experimental Subject is to record different people doing the same action, and the View is to record a viewpoint of different camera positions. We express the cross-performer and cross-view as (X-Sub) and (X-View), respectively. Then, we report the highest classification accuracy throughout the training process on X-Sub and X-View, respectively.

Table 3. Comparision performance of action recognition on the Kinetics dataset.

Methods	**Top-1 (%)**	**Top-5(%)**
RGB[21]	57%	77.3%
Feature Enc.[2]	14.9%	25.8%
Deep LSTM [20]	16.4%	35.3%
TemporalConv[1]	21.5%	42.6%
ST-GCN[14]	30.7%	52.8%
Our Scheme	32.8%	55.6%

Table 4. Comparision performance of action recognition on NTU-RGB+D dataset.

Methods	**X-Sub**	**X-View**
Deep LSTM [20]	60.7%	67.3%
TemporalConv [1]	74.3%	83.1%
ST-LSTM+TS [4]	69.2%	77.7%
C-CNN+MTLN[10]	79.6%	84.8%
ST-GCN [14]	81.5%	88.3%
Our Scheme	87.24%	96.94%

From Table 3 and Table 4, we can see that the proposed scheme has significant advantages in terms of recognition efficiency and accuracy. And our scheme ends up with 2.1% and 3.2% accuracy improvement on Kinetics for Top-1 and Top-5, respectively; meanwhile, our scheme ends up with 5.7% and 7.6% improvement over ST-GCN for X-Sub and X-View, respectively.

4.4. **Ablation Study.** In order to illustrate the validity of the DPST-GCN model in the pose recognition process, we need to conduct further additional ablation experiments. The ablation experiments are set up with (1) skeletal information as the input of the original ST-GCN model, (2) information of the large have orientation prediction as the input of the

original ST-GCN model, and (3) the model proposed in this paper. In terms of database selection, because the Kinetics dataset not only records human skeleton data, contains many background and irrelevant objects, which may lead to errors in pose estimation, detecting some objects as skeletons incorrectly. Therefore, we only request the NTU-RGB+D action recognition dataset as the experimental data for the ablation study.

TABLE 5. Experimental study of ablation of NTU-RGB+D action recognition dataset.

Methods	**X-Sub**	**X-View**
ST-GCN with bone stream	81.5%	88.3%
ST-GCN with bone direction stream	83.58%	89.5%
Our Scheme	87.24%	96.94%

In Table 5, we do the comparison experiments about different experimental data input configurations. The scheme based on ST-GCN with bone direction stream performs better than the bone stream, and we can see that the ablation experiments in NTU-RGB+D action recognition dataset prove the validity of the direction of the bones' motion.

5. **Conclusions.** In this paper, a new DPST-GCN method based on directional computation and prediction for skeleton-based action recognition is proposed. Our DPST-GCN model consists of key skeletal point detection, skeletal orientation generation prediction and spatio-temporal model training modules. Firstly,in the algorithm, we parameterize the graph structure of the skeleton data, abstracts the key skeletal nodes and inter-node relationships as vertices and edges of the graph structure, respectively, and builds a graph network structure with nodes and inter-node relationships. The generation and prediction of skeletal orientation information enriches the dimensionality of temporal and spatial information in ST-GCN. This data-driven approach not only increases the flexibility of the GCN, but also enables a more detailed representation of the spatio-temporal information of the human skeleton, which makes it more applicable to action recognition tasks.

Besides, based on the motion direction of the bone, two points with a connection relationship are taken to express the predicted information between two joints of the bone. The prediction yields nodes as new nodes, thus updating the graph network structure. Finally, all the generated new nodes form a new skeleton, which improves the information and discriminative power of the skeletal representation. Our proposed algorithmic model was evaluated on two large-scale action recognition datasets, NTU-RGBD and Kinetic, and achieved higher human discrimination efficiency in both cross-scene(X-Sub) and cross-view(X-View) experiments. In the future,we will further analyze the relationship among bone direction streams and the critical skeletal points to improve the model, and consider more contextual information, such as interactions, to aid action recognition.

Acknowledgment. This work was supported in part by the Natural Science Foundation of Zhejiang Provincial under Grant LQ20F030010 and in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under the Grant of GK199900299012-013.

REFERENCES

[1] T. S. Kim, A. Reiter, Interpretable 3D Human Action Analysis with Temporal Convolutional Networks, *Proc.IEEE Conference on Computer Vision and Pattern Recognition*, pp.1623-1631,2017.

[2] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, Modeling video evolution for action recognition,*Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.5378-5387,2015.
[3] A.Zhu, Q.Wu, R.Cui, et al. Exploring a rich spatial temporal dependent relational model for skeleton-based action recognition by bidirectional LSTM-CNN,*Neurocomputing*, vol.414, pp.90-100,2020.
[4] Q.She, G.Mu, H.Gan, et al. Spatio-temporal SRU with global context-aware attention for 3D human action recognition, *Multimedia Tools and Applications*, vol.79, no.17-1, pp.12349-12371, 2020.
[5] Y.Han, S. Chung,Q. Xiao, et al. Global Spatio-Temporal Attention for Action Recognition Based on 3D Human Skeleton Data.*IEEE Access*,vol. 8,pp.88604-88616,2020.
[6] J. Wu, C. Hu, and Y. Wang, A Hierarchical Recurrent Neural Network for Symbolic Melody Generation,*IEEE Transactions on Cybernetics*, vol.50, no.6, pp.2749–2757,2020.
[7] Y. Wu, L. Wei , and Y. Duan ,Deep spatiotemporal LSTM network with temporal pattern feature for 3D human action recognition,*Computational Intelligence*, vol.35, pp. 535–554, 2019.
[8] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao,Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5457–5466, 2018.
[9] H. Liu, J. Tu, and M. Liu, Two-Stream 3D Convolutional Neural Network for Skeleton-Based Action Recognition,*Computer Vision and Pattern Recognition*, vol.4, no.8, pp.1–5,2017.
[10] H. H.Pham, H.Salmane, et al. Spatio-Temporal Image Representation of 3D Skeletal Movements for View-Invariant Action Recognition with Deep Convolutional Neural Networks,*Sensors*,vol.19,no.8,pp.1932-1958,2019.
[11] H. Wang,B.Yu,K.Xia, et al. Skeleton edge motion networks for human action recognition,*Neurocomputing* vol.423,pp.1-12,2021.
[12] C. Li, Q. Zhong, D. Xie, and S. Pu, Skeleton-based Action Recognition with Convolutional Neural Networks, *IEEE Computer Society*, vol.24, no.5, pp. 624–628. 2017.
[13] S. Zhang, Y.Yang, J. Xiao, X. Liu, and Y.Yang, Fusing Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks,*IEEE Transactions on Multimedia*, vol.20, no.9, pp. 2330–2343,2018.
[14] S.Yan, Y. Xiong , D. Lin, Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,*Proc. Association for the Advance of Artificial Intelligence*,2018.
[15] L. Shi, Y. Zhang, J. Cheng, and H. Lu, Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition,*Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.12018–12027,2019.
[16] H. Tang, F. Ye,Skeleton-based action recognition with JRR-GCN,*Electronics Letters*, vol.55, no.17, pp. 933–935, 2019.
[17] D. Tian, Z.M. Lu, X. Chen, and L.H. Ma, An attentional spatial temporal graph convolutional network with co-occurrence feature learning for action recognition, *Multimedia Tools and Applications*, vol.79, no. 2, pp. 679–697, 2020.
[18] T. Ahmad, L.Jin , L. Lin , and G.Tang, Skeleton-based action recognition using sparse spatio-temporal GCN with edge effective resistance,*Neurocomputing*, vol.423, no.4, pp.389–398,2021.
[19] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis,*Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1010–1019,2016.
[20] W. Kay , J. Carreira, K. Simonyan , B. Zhang, and A. Zisserman, The Kinetics Human Action Video Dataset. 2017.
[21] R.Cui, A.Zhu, S.Zhang, and G.Hua, Multi-source Learning for Skeleton based Action Recognition Using Deep LSTM Networks. *Proc. International Conference on Pattern Recognition*, pp.547–552,2018.
[22] P.N.Huu, D.N. Tien, and K.N. Manh, Action recognition application using artificial intelligence for smart social surveillance system.*Journal of Information Hiding and Multimedia Signal Processing*,vol.13,no.1,pp.1-11,2022.