Research Article
Recognition of Hitting Action in Cyclic Anaerobic Volleyball by Acute Cooling Based on Improved Spatiotemporal Graph Convolutional Network

Jinwen Zhang and Tianxing Shen

1Ministry of Sports, Nanjing University of Information Science and Technology, Nanjing, 210014 Jiangsu, China
2Sports Institute, Nanchang University, Nanchang, 330031 Jiangxi, China

Correspondence should be addressed to Jinwen Zhang; 001331@nuist.edu.cn

Received 10 March 2022; Revised 25 April 2022; Accepted 5 May 2022; Published 26 May 2022

Copyright © 2022 Jinwen Zhang and Tianxing Shen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Action recognition is the basis of intelligent sports training and has a unique role in improving athletes’ sports training ability. Traditional motion recognition cannot accurately identify changes in human skeletal motion, and the generalization ability is weak, which cannot meet the needs of modern sports. This paper is aimed at studying a new method of human motion recognition and better realizing the intelligent development of the Institute of Physical Education. This paper proposes an improved recognition mode of spatiotemporal graph convolutional network, builds a human skeleton feature recognition model, learns the changes of human skeleton motion, and masters the laws of human skeleton motion changes. This paper comprehensively compares the differences of action recognition of several different algorithms and proves the superiority of this method. Afterwards, a comprehensive comparison of 2D and 3D information data and a comparison of whole body and local action recognition methods were conducted to investigate the experience of athletes and trainers. The experimental results of this paper show that the spatiotemporal graph convolutional network can better realize the recognition of motion features, improve the sports training ability of athletes by 20%, and promote the development of sports.

1. Introduction

In sports practice, the research on how to improve the effect of sports training has always been a frontier topic in sports research. From a macro perspective, motion is the constant change of the geometry and posture of an instrument of the human body. In a sense, the movement of the human body is the result of the orderly work of various joint muscles in different forms. The purpose of exercise technology analysis is to explore the relationship between exercise technology and the physical conditions that support it. One of the keys to using this relationship for effective training is to fully grasp the kinematic characteristics of each joint muscle group in special sports according to the structure and function of the human body, so as to formulate appropriate scientific training methods. Improving athlete’s athletic ability and technical level is critical to continuous improvement in athletic performance.

In order to reveal the kinematic characteristics of technical movements of volleyball players, athletes of different technical levels should follow the same rules when performing technical movements. However, due to the different development levels of athletes’ athletic qualities, the technical characteristics of athletes are also different. The technical characteristics of top volleyball players’ movements may represent the development trend of technical movements. With the continuous development of science and technology, visual observation by trainers has become a thing of the past, and instrument observation has become a modern tool for computer-aided analysis. The data obtained by people is more and more abundant and accurate, so that the research of analyzing sports technology has become
quantitative, objective, and scientific. Therefore, studying the kinematic characteristics of technical movements of high-level volleyball players under certain technical conditions is helpful to improve the training level of technical movements and enrich the theory of technical movements.

With the extensive research and rapid development of artificial intelligence technology, the recognition of human behavior has gradually entered people’s field. Human behavior contains a lot of behavioral information and is an important medium for human-computer interaction. In videos containing human motion, human skeleton is the main carrier of motion information. Compared with video frames, using skeleton to record motion data has the advantages of smaller storage space, illumination, and contrast protection. Therefore, the task of skeleton-based video action recognition has attracted the attention of researchers. One application of spatiotemporal graph convolutional neural networks is human skeleton recognition and detection. It extracts video data and converts it into skeleton data, which eliminates the error of action recognition and improves the accuracy of action recognition.

In this paper, we study an action recognition method based on fusion of human skeleton information and video images. Its advantages are as follows: (1) While maintaining the action relevance of the skeleton-based method, the accuracy of action recognition is improved by combining image information. (2) The influence of the background light interference factor is excluded, the action recognition is more accurate, and the error is smaller. (3) Combined with modern analysis technology, it can comprehensively analyze the movements, find the defects and deficiencies of the movements, and make targeted improvements and optimizations to help volleyball players better themselves.

2. Related Work

For the research on motion recognition, domestic and foreign experts and scholars have achieved certain results. Ding et al. designed a finger gesture recognition system based on mechanical myogram (MMG) to recognize the movement of each finger. SVM is a class of generalized linear classifiers that perform binary classification of data in a supervised learning manner. SVM can perform nonlinear classification through kernel methods and is one of the common kernel learning methods, which can be used for portrait recognition. The system uses support vector machine (SVM) to process three feature set classifiers: wavelet packet transform (WPT) coefficients, stationary wavelet transform (SWT) coefficients, and time-frequency domain hybrid (TFDH) features. The experimental results show that the average accuracy of recognition using WPT, SWT, and TFDH features is 91.64%, 94.31%, and 91.56%, respectively. Furthermore, when the three feature sets are combined in the proposed recognition system, the average recognition rate can reach 95.20% [1]. Anam et al. proposed a new index finger motion recognition system using a cutting-edge method and its improvement. The system he proposed consists of a combination of feature extraction methods, dimensionality reduction methods, and the well-known classifier support vector machine (SVM). An improvement to the SVM, self-advised SVM (SA-SVM) was tested to evaluate and compare its performance with the original SVM. Experimental results show that SA-SVM improves the classification performance by an average of 0.63% [2]. Li proposed a new method for mirror motion recognition of rehabilitation robots based on multi-channel sEMG signals. First, a muscle coordination model of basic sEMG signals was established using bilateral mirror training. Second, the constrained L1/2-NMF is used to extract the main sEMG signal information, which can also reduce limb motion features. Finally, the relationship between sEMG signal features and upper limb movements is described by TSSVD-ELM and applied to improve the stability of the model. The effectiveness and feasibility of the proposed strategy are verified by experiments, and the rehabilitation robot can move with the mirrored upper limbs. Comparing the proposed method with PCA and full motion feature extraction, it is confirmed that the convergence speed is faster, and the feature extraction accuracy is higher, which can be used for rehabilitation robotic systems [3]. Luo analyzed and discussed the human error action recognition method based on the idea of template matching, analyzed the key issues affecting the overall expression of error action sequences, and proposed a motion energy model based on direct motion energy decomposition. Video clips of human error actions in 3 Deron’s action sequence space are passed through a filter bank, which can avoid preprocessing operations such as object localization and segmentation. The MET feature highlights the harmony of human organs and structures. It is a compilation of body photography from all over the world. Then, the MET features are used in combination with SVM to test the human error database, and the experimental results obtained using different feature reduction and classification methods are compared. The results show that the method has obvious comparative advantages in recognition rate and is suitable for other dynamic scenes [4]. Kim et al. developed a low sampling rate multichannel surface electromyography (sEMG) module and applied it to hand movement studies. He studied the effect of sEMG signal sampling rate and feature extraction window length on the classification accuracy of hand action recognition. Ten normal subjects and one forearm amputee were asked to wear an armband module consisting of 8 EMG sensors to measure 7 and 4 hand movements of normal subjects and amputees, respectively. Hand motion recognition uses artificial neural networks, to support vector machines, decision trees, and k-nearest neighbor classifiers. The results show that the hand action classification accuracy increases with the sampling rate and window length [5]. Gao et al. develops and proposes a new encoding scheme based on a review of previous approaches using the dynamic map concept for human motion recognition in RGB-D over different modalities (depth, skeleton, or RGB-D data). The improved method generates efficient flow-guided dynamic graphs that can select high-motion windows and distinguish the order between motion small frames. The improved flow-guided dynamic map achieves state-of-the-art results on the large-scale Chalearn LAP IsoGD and NTU RGB+D datasets [6]. Camargo and Young evaluated feature selection in synchronized motion classifications generated by combinations of
wrist and elbow flexion/extension, radioulnar pronation/supination, and hand open/close, with the aim of identifying a general set of motion implementation recommendations from EMG signals. Classify for prosthetic control. Chow-Liu trees and forward feature selection are used as methods for selecting features, and six different classification algorithms are evaluated as wrapping components. Conclusion is that the Chow-Liu trees are shown to be a strategy that can match the accuracy achieved by forward selection search with a small number of iterations. Trends in waveform length and entropy were the most relevant feature types to consider and found evidence that synchronous motion classification should be handled using nonlinear classification methods [7]. Cho et al. proposed a hand motion recognition system based on forearm deformation. By using machine learning-based techniques, the proposed method can be applied to various users and various measurement conditions. First, an array of distance sensors was developed to measure forearm deformation. Then, the applicability of three machine learning-based classifiers (k-NN, SVM, and DNN) was tested and validated using the measured forearm deformations. In experiments, the accuracy of the proposed system is verified using different users. He also tested the system against different elbow postures and when measuring data on clothing [8].

3. Improved Spatiotemporal Graph Convolutional Neural Network and Principle of Motion Recognition

3.1. Introduction to Knowledge about Graphs. Graphs can describe relationships between data. Graphs can be classified into two categories according to the orientation of edges: undirected and directed graphs [9]. The human body structure basically exists symmetrically, and the adjacency matrix of an undirected graph must be a symmetric matrix, but not necessarily a directed graph. And the direction of human action is uncertain, so action recognition uses more undirected graphs. The human body structure is divided into two categories according to the orientation of edges: undirected and directed graphs [9]. The human body structure basically exists symmetrically, and the adjacency matrix of an undirected graph must be a symmetric matrix, but not necessarily a directed graph. And the direction of human action is uncertain, so action recognition uses more undirected graphs. The graphs used in this article are all undirected graphs, and Figure 1 is an example of an undirected graph.

Given an undirected graph $T = \{D, B\}$, where $D$ represents a set of vertices and $B$ a set of edges. The vertices of a set of vertices represent individuals in the real system, and the edges of a set of edges represent the relationships between individuals. The undirected graph $T$ contains a vertex $i$, and $x_m$ represents the vertex $m$ of the undirected graph $T$. The size of the adjacency matrix $I$ of the undirected graph $T$ is $i \times i$, and the number $J_{mp}$ of the $m$-th row $p$-th of the adjacency matrix $J$ represents the weight between the vertex $x_m$ and the vertex $x_p$. Since the undirected graph $T$ is an undirected graph $J_{mp} = J_{pm}$, the shape of the adjacency matrix is a symmetric matrix. When there is no connection between vertex $x_m$ and vertex $x_p$, setting the weight between them to 0. Letting vertex $x_m$ and its own weight $J_{mm}$ be 0. The definition of the degree $a_m$ of the vertex $x_m$ is as follows:

$$a_m = \sum_p J_{mp}. \quad (1)$$

The degree matrix $A$ of the undirected graph $T$ is a diagonal matrix, and the value on the diagonal is the degree of the vertex, which is defined as follows:

$$A = \text{diag}(a_1, a_2, \cdots, a_n). \quad (2)$$

A simple Laplace matrix $Q$ of an undirected graph $T$ is defined as follows:

$$Q = A - I. \quad (3)$$

The normalized Laplacian matrix $G$ in the undirected graph $T$ is defined as follows:

$$G = A^{-1/2}QA^{1/2} = R_m - A^{-1/2}JA^{-1/2}. \quad (4)$$

In the formula, $R_m$ represents an identity matrix of size $m \times m$. This paper mainly uses the normalized Laplacian matrix. The Laplace matrix is a semipositive definite matrix with $m$ linearly independent eigenvectors whose eigenvalues are not negative. The Laplacian matrix can be decomposed as follows:

$$Q = KIK^{-1} = KIK^V. \quad (5)$$

Among them, $K = (\tilde{k}_1, \tilde{k}_2, \cdots, \tilde{k}_n), \tilde{k}_1, \tilde{k}_2, \cdots, \tilde{k}_n$ is the Laplacian matrix $n$ eigenvectors, matrix $K$ is the orthogonal matrix, $I = \text{diag}(y_1, y_2, y_3, \cdots, y_n), y_1, y_2, y_3, \cdots, y_n$ is the Laplacian matrix $n$ eigenvalues, and $V$ is the inner product of the eigenvectors.

The Fourier transform is extended to graph theory [10]. The Fourier transform of a graph is defined as follows:

$$\tilde{h} = K^Vh. \quad (6)$$

In the formula, $h$ represents the $n$-dimensional vector formed by all vertex values in the undirected graph $T$, and $\tilde{h}$ is the $n$-dimensional vector after the Fourier transform in the undirected graph $T$. The inverse transform of the Fourier transform of a graph is defined as follows:

$$h = K\tilde{h}. \quad (7)$$

3.2. Graph Convolutional Neural Networks. The convolution operation is defined in the Fourier domain by computing the
eigendecomposition of the graph Laplacian operator. Fourier transform, which means that a function that meets certain conditions, can be expressed as a trigonometric function (sine and/or cosine function) or a linear combination of their integrals. The Fourier domain is the collection of Fourier-transformed signals in the instant domain. The spectral convolution on the graph can be defined as the product of the signal \( h \in \mathbb{F}^S \) and the filter \( d_a = \text{diag} (\alpha) (\alpha \in \mathbb{F}^S) \) in the Fourier domain, as follows:

\[
d_a * h = K d_a (Z) K^V h.
\]  

Chebyshev polynomials have important applications in approximation theory. This is because the roots of Chebyshev polynomials of the first kind (called Chebyshev nodes) can be used for polynomial interpolation. The corresponding interpolation polynomial minimizes the Runge phenomenon and provides the best consistent approximation of the polynomial over continuous functions. To simplify the computation, \( d_a (Z) \) can be approximated by a truncated expansion of the Chebyshev polynomial \( P_b (\tilde{Z}) \) up to the \( b \)-th order to get an approximation to \( d_a (Z) \):

\[
d_a (Z) = \sum_{b=0}^{b} \alpha^b P_b (\tilde{Z}),
\]

\[
\tilde{Z} = \frac{2}{Y_{\max}} Z - R_m.
\]  

The hierarchical eigenvector matrix for the largest eigenvalue of \( G \) (that is, the spectral radius) is \( \tilde{Z} \). A vector of Chebyshev coefficients can be denoted \( \alpha' \in \mathbb{F}^S \). Chebyshev polynomials are defined recursively [11]:

\[
P_b (m) = 2 m P_{b-1} (m) - P_{b-2} (m).
\]

Among them, \( P_0 (m) = 1, P_1 (m) = m \). Since it is a \( b \)-order polynomial in Laplace, it is \( b \)-localized. We use the approximate \( d_a' \) to replace the original \( d_a \); the formula is as follows:

\[
d_a' h = K \sum_{b=0}^{b} \alpha^b P_b (\tilde{Z}) K^V h = \sum_{b=0}^{b} \alpha^b K P_b (\tilde{Z}) K^V h.
\]  

Here, \( P_b (\tilde{Z}) \) is the \( b \)-order polynomial of \( \tilde{Z} \), with

\[
K \tilde{Z}^b K^V = (K \tilde{Z} K)^b = W^b,
\]

\[
W = \frac{2}{Y_{\max}} Z - R_m,
\]

\[
d_a' h = \sum_{b=0}^{b} \alpha^b P_b (W) h.
\]

Since the eigenvalues of \( R_m + A^{-1/2} Q A^{-1/2} \) range from [0, 2], stacking this operator may lead to numerical instability and exploding or vanishing gradients, and a renormalization trick is introduced [13]:

\[
R_m + A^{-1/2} Q A^{-1/2} \rightarrow \tilde{Q} = Q + R_m,
\]

\[
\tilde{A}_{mn} = \sum_n \tilde{A}_{mn}.
\]

Generalizing the definition to signal \( H \in \mathbb{F}^{E \times U} \) which has \( U \) input channels and \( E \) filters for feature mapping. That is, when the representation of each node in the graph is not a separate scalar but a vector of size \( U \), the formula is as follows:

\[
Y = A^{-1/2} Q A^{-1/2} H \zeta.
\]  

Hierarchical convolution operation is limited to \( B = 1 \) to alleviate the overfitting problem on the local neighborhood structure of graphs with very wide node degree distributions. At this point, the spectral convolution can be approximated as a linear function of \( W \). But this operation can only establish dependencies of first-order neighbors. For this problem, \( B \)-order neighbor dependencies can be established by stacking multilayer graph convolutional networks. There is no need to be restricted by Chebyshev polynomials when building dependencies of \( B > 1 \) order neighbors [12]. GCN is a very powerful neural network framework for machine learning on graphs. It is a neural network that operates on graphs and aims to directly process graphs and exploit their structural information. Further in the linear model of GCN, approximating \( Y_{\max} \approx 2 \) to obtain a first-order linear approximation of spectral convolution:

\[
d_a' h = \alpha (R_m + A^{-1/2} Q A^{-1/2}) h.
\]

Here, \( \alpha, \alpha' \) is a free parameter. If you need to establish \( B \)-order neighbor dependence, you can achieve it by setting the \( b \)-layer filter.

In order to avoid overfitting to constrain the parameters, letting \( \alpha = \alpha' = -\alpha' \) obtain the following formula:

\[
d_a' h = \alpha (R_m + A^{-1/2} Q A^{-1/2}) h.
\]

So far, it is found by approximation that the spectral graph convolution no longer depends on the entire graph, but only on the neighbor nodes within \( B \) steps from the central node. In order to reduce the computational cost, the hierarchical convolution operation is limited to \( B = 1 \) to alleviate the overfitting problem on the local neighborhood structure of graphs with very wide node degree distributions. At this point, the spectral convolution can be approximated as a linear function of \( W \). But this operation can only establish dependencies of first-order neighbors. For this problem, \( B \)-order neighbor dependencies can be established by stacking multilayer graph convolutional networks. There is no need to be restricted by Chebyshev polynomials when building dependencies of \( B > 1 \) order neighbors [12]. GCN is a very powerful neural network framework for machine learning on graphs. It is a neural network that operates on graphs and aims to directly process graphs and exploit their structural information. Further in the linear model of GCN, approximating \( Y_{\max} \approx 2 \) to obtain a first-order linear approximation of spectral convolution:

\[
d_a' h = \alpha (R_m + A^{-1/2} Q A^{-1/2}) h.
\]  

Since the eigenvalues of \( R_m + A^{-1/2} Q A^{-1/2} \) range from [0, 2], stacking this operator may lead to numerical instability and exploding or vanishing gradients, and a renormalization trick is introduced [13]:

\[
R_m + A^{-1/2} Q A^{-1/2} \rightarrow \tilde{Q} = Q + R_m,
\]

\[
\tilde{A}_{mn} = \sum_n \tilde{A}_{mn}.
\]

Generalizing the definition to signal \( H \in \mathbb{F}^{E \times U} \) which has \( U \) input channels and \( E \) filters for feature mapping. That is, when the representation of each node in the graph is not a separate scalar but a vector of size \( U \), the formula is as follows:

\[
Y = A^{-1/2} Q A^{-1/2} H \zeta.
\]  

Among them, \( \zeta \in \mathbb{F}^{E \times O} \) is expressed as a parameter matrix, and \( Y \in \mathbb{F}^{U \times O} \) is expressed as a convolution signal matrix, that is, the convolution result. At this time, the node representation of each node is updated into a new \( O \)-dimensional vector, which contains the information on the corresponding first-order neighbors. This leads to the final form of the graph convolutional neural network:

\[
S^{(i+1)} = \delta \left( \tilde{A}^{-1/2} Q A^{-1/2} S^{(i)} M^{(i)} \right).
\]  

(17)
The input of the \( t \)-layer network is \( S^{(t)} \in F^{U \times L} \), the initial input signal is \( S^{(0)} = H \), \( U \) is the number of nodes in the graph, and each node is represented by an \( L \)-dimensional feature vector. \( \tilde{Q} = Q + R_m \) is expressed as additive self-connected adjacency matrix, \( \tilde{A} \) degree matrix, \( \tilde{A}_{mn} = \sum_{n} \hat{A}_{mn} \). \( M^{(t)} \in F^{L \times L} \) is the weight parameter to be trained. \( \delta \) is the activation function.

The skeleton graph data designed for video human action recognition is dynamic and can reflect the real dependencies between nodes according to the predefined graph structure. In some cases, learning the underlying dynamic spatial correlation can further improve the accuracy of the model. The human skeleton structure graph contains rich motion information, so graph convolution can be considered to aggregate the features of different nodes in different time frames.

### 3.3. Skeleton Spatiotemporal Feature Extraction Based on Graph Convolutional Network

Current motion recognition methods can be divided into two categories: motion recognition methods based on video data and motion recognition methods based on skeleton data. In general, video data is easy to collect, but recognition methods based on video data must overcome interference from background, lighting, and occlusion. The identification method based on skeleton data can effectively avoid the interference of these factors, but the skeletal data is not easy to collect [14]. This article discusses the necessity of using video data for technical volleyball sports. Combined with experimental analysis, this paper processes the action video data of batting volleyball technology, converts it into skeleton data, and inputs it into the action recognition network.

This paper uses the method of human pose estimation to preprocess the video data. Human pose prediction is a very basic problem in computer vision. Simply putting, it is the position estimation of the posture of the “human body” (key points are such as head, left hand, and right foot). It has two ideas; one is “top-down,” which means that the human body area is detected first, and then, the key points of the human body in the area are detected. The second is “bottom-up,” which means to first detect all the key points of the human body in the picture and then map these key points to different individuals. For video data containing human body information, the human body pose prediction can be used to fit the two-dimensional or three-dimensional coordinates of the human skeleton and joints in the video through the related methods of deep learning. The human pose prediction method used in this paper can be divided into three steps [15].

**Step 1.** Using a convolutional neural network to detect the position of the human body in each frame of the input video data and using a convolutional neural network (such as VGG16) can output the feature map of the picture. By training the parameters for the labels of the given data, the network will have a higher activation value for the pixels in the area where the human body is located. The trained feature map is remapped to the original image size by upsampling, and multiple bounding boxes can be output, and each bounding box represents the possible position of the human body.

**Step 2.** The pose of the human body in each bounding box is predicted, the pose of each frame is filtered, and the 2D coordinates of the human joints are estimated [16]. Different from training the network with the bounding box of the position of the human body as the label in the human body detection work, the training label of the human body pose estimation is the position of the human body joint point in the picture. The method of outputting the coordinate positions of human joint points after network training is usually called single person pose prediction (SPPE). It is not suitable to directly input the picture in the bounding box output by the human position detection as the pose prediction network. Before inputting the SPPE, the data is preprocessed by a layer of spatial transformation network (STN). Spatial transform network (STN) is a convolutional neural network architecture model. By transforming the input image, the influence of the spatial diversity of the data is reduced, and the classification accuracy of the convolutional network model is improved. The robustness of STN is very good, and it has spatial invariance such as translation, expansion, rotation, perturbation, and bending. The working principle of STN is roughly: first is generating spatial transformation parameters through a series of network layers (such as fully connected or convolutional neural networks) to achieve two-dimensional affine transformation. Then, the input image is resampled through this parameter to realize the translation, scaling, or rotation of the image data, which enhances the robustness of the subsequent network. The image in the preprocessed bounding box is input into SPPE, and each bounding box will generate a pose estimation of a set of human joint points, and each joint point will have its corresponding confidence. For all joints corresponding to each human joint position, only one joint with the highest confidence is retained, and this step is called nonmaximum suppression (NMS). The output joint points are connected to complete the conversion of image data to two-dimensional human body joint point data.

**Step 3.** Using the two-dimensional coordinates of the human body joints as input, fitting the three-dimensional coordinates of the human body joints is shown in Figure 2. Using the coordinates of the joint points of the human body in the three-dimensional space as labels, the two-dimensional coordinates can be converted into three-dimensional coordinates through the network.

Human detection on initial video data uses the VOC_2007 dataset [17] to train network parameters. The dataset has a total of 9963 images, of which 5011 are used for training and 4952 are used for testing. Each image is marked with a bounding box where the body is located. Human pose prediction by converting video image data to skeleton three-point joint point data uses the Human 3.6M dataset [18] to train network parameters. The dataset has 3.6 million images, and each image is labeled with the 3D coordinates of human joint points. There is a total of 17 subjects (9 males and 8 females, generally choosing 2, 3, 6, 10, and 12 as
training data and 13 and 14 as test data), a total of 24 action scenes, such as serving, hitting, jumping, and throwing. This data is captured by 5 digital cameras, 2 time sensors, and 12 motion cameras. Both human detection and human posture prediction belong to the research of regression problem, so the network parameters obtained by training with other datasets still have good adaptability to the volleyball hitting action dataset used in this paper.

The human skeletal structure can be regarded as a structure formed by the connection of joints and bones. In this paper, a graph structure model with joint points as graph nodes and bones as edges is established. The coordinate position information of the joint points is input into the network as the feature of the graph nodes, and the spatial structure information of the joint points is processed by the graph convolutional network (GCN). The features in the temporal dimension are processed by temporal convolution (TCN) to generate high-level feature maps [19]. TCN is a time series convolutional neural network model, a new type of algorithm that can be used to solve time series forecasting. TCN provides a unified approach to capture all two levels of information hierarchically. Finally, the corresponding action category is performed by the SoftMax classifier.

For the human joint point sequence generated by the human pose prediction network, an undirected graph constructs the human skeleton structure, in which the nodes connect the skeleton sequence, and the joint fusion nodes constitute the human skeleton structure. The human skeleton can be simplified into a structure composed of 9 joint points, so the output model of the human motion recognition system can be represented as Figure 3.

Defining acquisition functions at different joints to ensure that the samples sampled each time are the same. If the sampled samples are not the same, add the empty set data as a node on the sample to achieve the consistency of the sample data. The convolution kernel is given the input image during image processing. The weighted average of pixels in a small area in the input image becomes each corresponding pixel in the output image, where the weights are defined by a function called the convolution kernel. In this paper, a convolution kernel is designed for joint connection in time dimension. On a single-frame human joint-point connection structure, a sampling function is defined on the joint point’s neighbor set. At the same time, it is ensured that each sampling is sampled. Its sampling function time series is certain, and the moving direction is shown in Figure 4. It can convolve all time series to ensure that the data information of the model is rich enough.

3.4. Graph Spatiotemporal Feature Extraction Based on Adjacency Matrix. Laplacian matrix is also called admittance matrix, Kirchhoff matrix, or discrete Laplacian operator. It is mainly used in graph theory, as a matrix representation of a graph. It can be used for the calculation of graphics three dimensions and can represent complex geometric structures. Laplacian matrix is an excellent method for processing graph information due to its positive definiteness [21], so Chebyshev convolution based on Laplacian spectral decomposition is used to extract the non-Euclidean correlation of joint points. At the same time, the parameter K is used to distinguish the features of adjacent nodes of different orders. First, the Laplacian matrix corresponding to each subgraph can be obtained through the adjacency matrix. Although joints move in groups when people perform an action, a joint may appear in multiple different positions relative to the human torso due to different actions performed. Therefore, each joint should be given different importance when modeling the dynamics of different joints for different actions. For example, when playing football, the movement of the legs may be more important than the neck. By looking at the movement of the legs, we can even tell which state the athlete is running, walking, and jumping in. However, in these states, the displacement changes of the neck may not contain much effective information worthy of our reference. In this sense, the Laplacian matrix corresponding to the
adjacency matrix representing the side information is multiplied by a learnable parameter mask $Y$ at each layer of the spatial graph convolution. It can effectively deal with the importance of different skeleton edges in the network modeling process [22–25].

Before the network is established, the feature entries of the graph, namely, the character dimension $M$ and batch dimension $B$ of the five-dimensional matrix $X$, and the feature dimensions of node dimension $N$ and position $C$ are merged. Then, the combined matrix is sent to one-dimensional batch normalization to obtain the input normalized matrix. Although this operation introduces parameters, the advantages outweigh the disadvantages. First, the joint positions of the same joint vary greatly from one frame to another. Secondly, after normalization, the joint positions under different frames in different data batches basically obey random distribution, which will not cause large differences in video data and lead to accuracy fluctuations. These normalized batch data matrices are split into node dimension and location feature dimension, which are denoted by $N$ and $C$, respectively. This matrix is input as $h^0$ into a two-dimensional convolutional layer with a convolution number of $3\times$kernel-num, a convolution kernel size of kernel-size, and a stride of $s$. Feature extraction is performed on the temporal dimension and node dimension of the spatiotemporal skeleton video graph. For the convenience of description, the number of convolution kernels is referred to as $k_n$, and the size of convolution kernels is $k_s$ in the following content. The obtained characteristic matrix is reprocessed, and finally the Chebyshev polynomial matrix is obtained. In order to comprehensively consider the features on each subgraph, feature averaging and dimension transformation are performed on the three subgraphs. Finally, the new Chebyshev polynomial matrix is obtained and multiplied with the learnable parameter matrix. In order to prevent the gradient from disappearing due to the network being too deep, when the number of input and output feature maps is the same, the input features are superimposed into the output of the learnable parameter matrix element by element, and the output feature $h^l$ of the first layer is obtained.

The network frame diagram of the first layer of the final spatial feature extraction network is shown in Figure 6. The output result of the last layer is input into the average pooling layer, and finally the output logical value is obtained. Then, the batch dimension $B$ and the character dimension

![Figure 4: Schematic diagram of the processing method of information in the time dimension.](image)

![Figure 5: Human action recognition network structure under the framework of spatiotemporal graph convolution.](image)
M in the logical value are split, and the features of different skeletons are voted on average. Finally, it is sent to the convolutional network to obtain the final output category prediction matrix, and the prediction matrix is compared with the real category through the cross-entropy loss function. The cross-entropy loss function is often used in classification problems, especially when neural networks do classification problems; cross-entropy is often used as a loss function. Also, since cross-entropy involves calculating the probability of each class, cross-entropy appears with the sigmoid (or softmax) function almost every time.

3.5. Volleyball Hitting Technique. In volleyball, the batting technique, also known as the hitting technique, refers to the method in which the batter jumps up and hits the ball into the opponent’s field net. This is also the main technique of volleyball. From the 1950s to the present, the development process of more than 60 years shows that with the rapid development of volleyball, the racquet technology has been greatly improved and innovated. The offensive form

---

**Table 1:** Comparison of input and output dimensions of spatiotemporal graph convolutional networks.

<table>
<thead>
<tr>
<th>Number of network layers</th>
<th>Input dimension</th>
<th>Output dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>First floor</td>
<td>(9,3,300)</td>
<td>(9,64,300)</td>
</tr>
<tr>
<td>Second floor</td>
<td>(9,64,300)</td>
<td>(9,64,300)</td>
</tr>
<tr>
<td>Third floor</td>
<td>(9,64,300)</td>
<td>(9,128,150)</td>
</tr>
<tr>
<td>Fourth floor</td>
<td>(9,128,150)</td>
<td>(9,128,150)</td>
</tr>
<tr>
<td>Fifth floor</td>
<td>(9,128,150)</td>
<td>(9,256,75)</td>
</tr>
<tr>
<td>Sixth floor</td>
<td>(9,256,75)</td>
<td>(9,256,75)</td>
</tr>
<tr>
<td>Seventh floor</td>
<td>(9,256,75)</td>
<td>(9,256,75)</td>
</tr>
<tr>
<td>Eighth floor</td>
<td>(9,256,75)</td>
<td>(9,256,75)</td>
</tr>
<tr>
<td>Ninth floor</td>
<td>(9,256,75)</td>
<td>(9,256,75)</td>
</tr>
</tbody>
</table>

**Table 2:** Comparison of the recognition method in this paper and other methods based on the self-built volleyball hitting action database.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSN</td>
<td>60.28%</td>
</tr>
<tr>
<td>Attention is all we need</td>
<td>67.21%</td>
</tr>
<tr>
<td>ST-GCN (enter the two-dimensional coordinates of the joint point)</td>
<td>69.62%</td>
</tr>
<tr>
<td>ST-GCN (enter the three-dimensional coordinates of the joint point)</td>
<td>71.17%</td>
</tr>
<tr>
<td>The method of this paper (input the two-dimensional coordinates of the joint point)</td>
<td>84.35%</td>
</tr>
<tr>
<td>The method of this paper (input the three-dimensional coordinates of the joint point)</td>
<td>92.89%</td>
</tr>
</tbody>
</table>
of volleyball has gradually developed from the initial hitting and fast attacking form to the three-dimensional attacking form that makes full use of the net and depth. In volleyball, spiking is the most effective way to score, and a team’s spiking ability is its offensive performance. At the same time, a strong spike is a way for the team to get rid of the passive and strive for the initiative, and it is the key to winning the game.

Volleyball batting techniques are divided into five types according to the action methods: As shown in Figure 7, they can be divided into frontal slamming, small swing arm slamming, one-foot jumping, and hooking arm slamming. Among them, the frontal punching technique is the foundation. The time of volleyball matches determines that volleyball players mainly rely on aerobic energy. During the competition, athletes show amazing bouncing power and fast arm swings. They also need athletes to start, brake, move, change direction, turn, block, and smash the ball. These all require athletes to have strong anaerobic capacity. The volleyball spiking action is only related to the state of the athlete and has nothing to do with the temperature. Therefore, when studying the identification of volleyball, we only need to consider the state of the athlete itself, without considering the influence of other factors such as rate and temperature.

4. Volleyball Movement Recognition Experiment and Analysis

4.1. Volleyball Sports Data. For the basic volleyball action dataset, convert the video data to skeletal data. The three-dimensional information of nine local joints is generated by the method of human pose prediction. Finally, the output of the network is constructed sequentially. For any initial skeleton dataset input to the network, the dimension is (9, 3300), which represents the number of joint points of the human body and the information dimension of the input joint points, usually the dimension of the coordinates. The number of joint points is three-dimensional, that is, three-dimensional space coordinates, representing the total number of frames of the input video. In this paper, a time series of data is populated to ensure a total frame count of 500, and this series is used as the input to the human motion recognition network. The 9-layer GCN-TCN module can be divided into three parts. The first three layers define the number of feature channels and the output of a single node as 64, the middle three layers are 128, and the last three layers are 256. The input and output dimensions are shown in Table 1. Dropout processing is used in each spatiotemporal convolution module. The final output is global pooling (GAP) and archived using softmax. Overall, using the standard cross-entropy loss function, the initial
learning rate is set to 0.0075 and reduced to 10% every 10 epoch iterations, resulting in 80 epochs. The experiment was tested and trained using an Intel E5-2630v4 platform and 4 GeForce RTX 2080 ti.

To verify the effectiveness of the proposed volleyball technique action recognition method, the training set of the volleyball technique action dataset I is used to train the network model, and the test set is used to test the accuracy of the model. Table 2 shows the use of dataset I.

Experiments show that the recognition rate of using the spatial three-dimensional coordinate information of local joint points as input is 92.89%, which is better than the recognition rate of 84.35% using the plane two-dimensional coordinate information of local joint points as input. Compared with other methods, the recognition rate of the method proposed in this paper is significantly higher.

In order to better test our system model in the field, randomly select a group of players to participate in closed volleyball sports for 7 days, shoot their batting mobilization training on the spot, and analyze the movement trajectory of the players, so that the trainer can guide the training of the players. During training, the evaluation of the motion recognition system by athletes and trainers was collected.

4.2. Data Analysis. As shown in Figure 8, in the result of using the spatial three-dimensional coordinate information of the local joint points as the input, the recognition accuracy of the hook smash and the small arm smash both reached more than 98.87%. However, there are similarities between frontal spiking and single-leg take-off spiking, and the degree of distinction is not high.

As shown in Figure 9, the accuracy of local joint point recognition is significantly higher than that of whole body joint point recognition. The reason is largely because different athletes have different exercise habits, and athletes may have invalid movements during exercise, so the accuracy of whole body recognition will be reduced due to invalid movements. Because the influencing factors are small, the local identification is less disturbed, and the recognition accuracy is higher.
As shown in Figure 10, by comparing the two-dimensional and three-dimensional comparison maps of local joint point information, it can be found that the recognition rate of the two-dimensional image is significantly lower than that of the three-dimensional image. Especially in the initiation part of the action, lack of one dimension of information, the recognition rate of the action will drop significantly, which is consistent with the results of other experiments.

As shown in Figure 11, it can be seen from the figure that with the extension of time, the satisfaction of trainers and players with the volleyball motion recognition system gradually increases. The possible reason is that with their experiments, the motion recognition system can well guide volleyball teaching training, help players correct their movements, and grow faster. Among them, it cannot be ignored that the convenience of the motion recognition system is not high, which is also the direction of future improvement.

5. Discussion

The article compares the recognition rate of this method and other methods for volleyball movement. It is obvious that the recognition method of this paper is taught by other methods, showing a good effect of volleyball hitting movement recognition. Further comparing the differences between 2D recognition and 3D recognition in this paper, the results of the experiment are in line with our cognition, and the recognition rate of adding one dimension is significantly higher. Comparing the whole body joint point recognition and the local joint point recognition again, it can be seen that the local recognition is more accurate and the accuracy is better. After investigating the feelings of athletes and trainers, we confirmed the superiority of this system. It can guide exercise practice very well, but there is also a certain problem of low convenience. Overall, the motion recognition system in this paper can guide the actual volleyball training very well and has strong practicability. Aiming at the attraction of motion recognition, the advantages of the method in this paper will be further strengthened, and the lack of convenience of the motion recognition system will be further improved, so that the motion recognition system can better serve athletes and trainers and improve the quality of sports.

6. Conclusion

In this paper, we start with the recognition of volleyball hitting action and convert video data into skeleton data with the help of spatiotemporal graph convolutional network, so as to better measure the movement status of athletes and improve the accuracy of action recognition. It has certain guiding significance for athletes’ anaerobic exercise training. The article fully introduces the principle of the spatiotemporal graph convolutional network and also takes you to understand the action charm of volleyball hitting, which has certain practical significance, but the article also has shortcomings. The existing human skeleton diagrams are artificially set according to the articulation relationship between the skeleton joint points and the bones, which do not conform to all action representations. Volleyball is a complex sport with various technical movements. The volleyball technical action data set built in this paper is not rich enough. The information fusion method used in this paper is relatively primitive. In the future, the motion recognition model should further enrich technical actions, optimize technical algorithms, and improve the accuracy of action recognition. At the same time, it is necessary to classify volleyball movements more accurately, so that they conform to human movements and better express human movements.

Data Availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Conflicts of Interest

The authors state that this article has no conflict of interest.

References


