International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1986-1999

https://doi.org/10.70517/ijhsa464171

Exploring the Application of Artificial Intelligence Technology in the Intelligent Management of Sports Education and Teaching Content

Xiaoyan Zhang[1,*]
[1] College of Physical Education and Health, Anhui Vocational and Technical University, Hefei, Anhui, 230001, China
Corresponding authors: (e-mail: zxy_uta@163.com).

Abstract The rapid development of artificial intelligence in the field of education has injected new momentum into the systematic transformation of educational models, driving innovation and optimization in teaching methods, evaluation systems, and management models. This paper uses the OpenPose algorithm to obtain skeletal point data for sports movements and standardizes the skeletal coordinate data. Based on the ST-GCN model, a multi-scale temporal attention mechanism is introduced to enhance the model's feature extraction capabilities, and a residual module is incorporated into the GCN to improve the model's local feature extraction performance. On this basis, the DTW algorithm is used to construct a sports movement evaluation model. Experiments show that the improved ST-GCN model achieves a MAP value of 79.3%, which is 9% and 6.8% higher than the MAP values of the image-based pose estimation algorithms SimpleBaseline and HRNet, respectively. The overall score for sports actions based on the DTW algorithm is 88.19 points, differing from the manually scored results by only 1.12 points. Integrating artificial intelligence technology with sports education and teaching can significantly enhance the intelligent reform of sports education and provide a technological foundation for the personalized development of sports education.

Index Terms artificial intelligence technology, OpenPose, attention mechanism, ST-GCN model, DTW algorithm

I. Introduction

As global educational reforms continue to deepen, teachers' classroom teaching practices are undergoing significant transformations [1]. In traditional physical education classroom teaching models, teachers primarily employ “lecture-based” and “rote learning” teaching methods, which emphasize the transmission of knowledge and skills, primarily achieved through explanations and demonstrations, while interactive activities, questioning, and discussion in the classroom are often neglected. Under the broader context of educational reform, innovative teaching models such as “inquiry-based learning,” “collaborative learning,” “project-based learning,” and “contextual learning” have been integrated into physical education classroom practices [2], [3]. Consequently, the objectives and forms of teaching behavior have undergone significant changes. The purpose of teaching behavior has shifted from the acquisition of knowledge and skills to the comprehensive development of students' various abilities, fostering their personalized development, and actively cultivating their autonomous learning capabilities [4]-[7]. The forms of teaching behavior have shown a trend toward diversification, with a significant increase in interactive, questioning, and discussion activities in the classroom, placing greater emphasis on the guiding role of teaching behavior in student learning [8]. Against the backdrop of rapid information technology development, artificial intelligence (AI) technology, with its powerful data processing and intelligent analysis capabilities, has gradually penetrated the education sector, significantly driving the transformation of educational models and teaching methods [9]-[11].

In their study on the application of artificial intelligence technology in sports teaching resource management, Lee, H, and Lee, J [12] examined the potential transformative role of artificial intelligence in sports education through customized personalized courses, knowledge transmission, assessment, and consultation, while emphasizing that future sports teachers need to enhance their professional capabilities in artificial intelligence. Wang, C, and Wang, D [13] proposed utilizing cloud computing, virtualization technology, and AI technology to manage physical education teaching resources in Chinese universities. Their research found that compared to traditional methods, this approach could improve resource sharing and utilization rates, thereby promoting the development of higher education. Hu, Z et al.[14] explored the trend toward AI-driven intelligent transformation in physical education and future research directions, identifying current limitations such as high technical costs, insufficient teaching resources,

and data security issues, to provide a framework for future development. Li, S et al.[15] developed a multi-feature fuzzy evaluation model based on AI technology to address challenges in assessing physical education teaching methods in universities. Liu, T et al. [16] proposed an AI-optimized physical education activity development model aimed at enhancing students' physical fitness and motor skills under the new curriculum reform. This model achieves high accuracy in motion recognition and meets practical needs. In summary, the application of AI technology in physical education resource management has significantly improved resource utilization efficiency [17]. Through intelligent management and analysis, teachers can quickly access and process large amounts of teaching resources, reducing the time and effort required for traditional manual organization. For example, artificial intelligence can be used to collect and analyze students' movement data in real time. The accumulation and analysis of this data provide a solid foundation for teachers to develop personalized teaching plans [18].

The introduction of artificial intelligence technology has injected new possibilities into physical education teaching, from scientific movement skill training to personalized exercise plan development, providing more efficient and precise tool support for physical education. The article relies on the OpenPose algorithm to obtain skeletal point coordinate data and standardizes the data. By combining the ST-GCN model with the attention mechanism, an improved ST-GCN model based on a multi-scale temporal attention mechanism was constructed for human motion recognition and estimation in physical education teaching. Based on the motion recognition results, a physical education motion evaluation model was constructed using the DTW algorithm. Through simulation experiments, the intelligent application effects of artificial intelligence technology in physical education teaching were verified.

II. Skeletal point data acquisition and preprocessing based on OpenPose

Artificial intelligence is having a revolutionary impact on the development of education, and physical education, as an important part of education, should also actively embrace artificial intelligence so that it can play a role in promoting physical education in the new era. Therefore, when faced with students who have grown up in the age of artificial intelligence and will be influenced by it throughout their lives, physical education has a strong need for the application of artificial intelligence.

II. A. Research on Pose Matching Based on OpenPose

II. A. 1) Overview of OpenPose

The OpenPose algorithm represents a significant advancement in the field of computer vision, particularly in the realm of real-time multi-person 2D pose estimation. This algorithm employs PAF detection to identify human body parts in images or videos. It estimates poses by iteratively refining partial affinity fields and partial detection confidence maps through consecutive stages. The architecture begins with a feedforward network, predicting a 2D confidence map of body part locations and a 2D vector field of partial affinity fields, encoding the degree of association between body parts. This setup enables OpenPose to process multiple individuals in an image, identifying anatomical keypoints for each person. By optimizing the network framework, larger convolutional kernels are replaced with multiple smaller ones to reduce computational load while maintaining accuracy. This approach combines deep learning techniques with iterative refinement of pose estimation.

OpenPose has been widely recognized for its performance and efficiency, winning the COCO Keypoint Challenge. The COCO Challenge is one of the most influential image recognition challenges in the field of artificial intelligence, aimed at advancing research and development in computer vision. The COCO dataset, proposed by Microsoft Research, is a large-scale computer vision dataset commonly used for analyzing and evaluating common visual tasks, including object detection, instance segmentation, human keypoint detection, and panoramic segmentation.

II. A. 2) OpenPose Principle

OpenPose is an open-source library developed based on convolutional neural networks and supervised learning, which implements pose estimation for human movements, finger movements, and facial expressions. It demonstrates excellent robustness in pose estimation under both single-person and multi-person conditions, thanks to its use of a bottom-up human pose estimation algorithm that utilizes human joint vector fields (PAFs) [19]. The process of OpenPose for input image processing is as follows:

First, calculate the joint confidence map and joint vector field to identify specific human joint points. Then, use the binary matching algorithm to connect the corresponding joint points for each person, thereby forming a human skeleton. OpenPose takes an image of size $w \times h$ as input and outputs the two-dimensional positions of the skeleton joint points for each person in the image.

The feedforward network outputs a set of human joint confidence maps S and a set of human body part vector fields L. Where $S = (S_1, S_2, \ldots, S_J)$ and $S_j \in R^{n \times h}$, $j \in [1 \ldots J]$, where J is the number of human joint nodes. $L = (L_1, L_2, \ldots, L_C)$ where $L_c \in R^{w \times h \times 2}$ and $c \in [1 \ldots C]$, with C being the number of body parts.

The feedforward network model architecture consists of two parts: the front half is a pre-trained backbone network, which is VGG-19 in OpenPose, and the back half uses two branches to output joint confidence maps and joint vector fields, respectively.

The two branches go through k stages to predict the output joint confidence maps and joint vector fields. First, the input image is processed by the first part of the model to obtain a feature map F. In the first stage, the feature map F is fed into the two branches of the first stage, which output corresponding results: the joint confidence map set S and the body part relationship field set L, where S contains J confidence maps and L contains C body part relationship fields.

The outputs of the first stage and the subsequent $k-1$ stages can be represented as:

$$S^1 = \rho^1(F) \tag{1}$$

$$L^1 = \phi^1(F) \tag{2}$$

$$S^t = \rho^t(F, S^{t-1}, L^{t-1}), \forall t \geq 2 \tag{3}$$

$$L^t = \phi^t(F, S^{t-1}, L^{t-1}), \forall t \geq 2 \tag{4}$$

After each stage is completed, the output results are merged, and the original feature map F is concatenated as the input for the next stage. Then, the Hungarian algorithm is used to perform maximum bipartite graph matching. After matching all adjacent joints, the joints of each person are obtained, and the optimal connection results for multiple joints are finally obtained.

II. B.Acquisition and standardization of skeletal point coordinate data

II. B. 1) Acquisition of skeletal point coordinate data

Joint points can describe human posture more directly and clearly, and human movement characteristics can also be represented by the movement of joint points. In dynamic environments, to reduce experimental costs, joint node positions are estimated from human motion images using human pose estimation technology. The joint node names required for the experiment are: left shoulder, right shoulder, neck, spine, hips, left elbow, right elbow, left hand, right hand, left knee, right knee, left foot, right foot, and head, numbered K1 to K14. The coordinates of the hip joint are approximately estimated using the intersection point of the straight lines connecting the right shoulder and left knee with the straight lines connecting the left shoulder and right knee. To ensure that the movement trajectories of joint points remain within the same range across different actions, each joint point is centered. Among these, the centering of all joint points $P_i(x_i, y_i, z_i)$ around the hip can be expressed as:

$$P_i(x_i, y_i, z_i) = P_i(x_i, y_i, z_i) - P_5(x_5, y_5, z_5) \tag{5}$$

After centralizing the joints, a simple skeleton is formed. The simple skeleton undergoes preprocessing, which includes skeleton scale normalization, skeleton perspective rotation, and removal of invalid frames. This paper only performs scale normalization on the skeleton, with the hip joint's position as the origin of the coordinate system. The vector formed by the spine and hip joint is set to $P_4(x_4, y_4, z_4)$, and the scale normalization standard ensures that the magnitude of the vector formed by the spine and hip joint is 1. Similarly, other joints are transformed in the same manner:

$$P_i(x_i, y_i, z_i) = P_i(x_i, y_i, z_i) / |P_4(x_4, y_4, z_4)| \tag{6}$$

II. B. 2) Skeletal coordinate standardization processing

Since coordinate point data is based on pixels, the skeletal point coordinate data obtained from the same video frame image in videos with different resolutions will differ. Therefore, the obtained skeletal coordinate data must be standardized and scaled [20].

Define the coordinate data of the i nd skeletal point at j resolution as $p_{i,j} = (x_{i,j}, y_{i,j}, c_{i,j})$. For example, taking the Nose with joint number 0 as an example, the coordinate data at 280p and 720p resolutions are:

$$\begin{aligned} p_{0,280p} &= (136.911, 216.846, 0.857697) \\ p_{0,720p} &= (352.442, 558.547, 0.855352) \end{aligned} \tag{7}$$

We measure the difference in bone point coordinates at different resolutions using point scaling ratios, i.e.:

$$k_p = \frac{x_{i,280p}}{x_{i,720p}} = \frac{y_{i,280p}}{y_{i,720p}} \tag{8}$$

The shoulder joint coordinate data for numbers 2 and 5 are as follows:

$$\begin{aligned} p_{2,280p} &= (115.510, 248.001, 0.908061) \\ p_{5,280p} &= (162.244, 244.141, 0.923461) \\ p_{2,720p} &= (297.227, 638.978, 0.906463) \\ p_{5,720p} &= (417.859, 629.000, 0.932781) \end{aligned} \tag{9}$$

Then, using the point-to-point distance formula $l = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, we can calculate the length of the shoulder joint.

Similarly, we measure the length of the limbs at different resolutions using the limb scaling ratio, i.e.:

$$k_l = \frac{l_{280p}}{l_{720p}} \tag{10}$$

The skeletal point coordinate data at different resolutions, point coordinate data, and limb lengths have the same scaling ratio, so we can directly use the position data of the Nose skeletal point with index 0 in the first frame of the video at two resolutions to obtain the scaling ratio of the limb length.

Therefore, joint point standardization is performed to achieve consistent pixel scale, and the transformation is obtained using the following equation.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} z & 0 \\ 0 & z \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{11}$$

where (x', y') are the coordinates of the human skeleton points in the normalized image, (x, y) are the coordinates of the human skeleton points in the source image, $z = x_{o,j1} / x_{0,j2}$, z is the scaling ratio, $x_{o,j1}$ is the coordinate data of the nose in the normalized image, and $x_{0,j2}$ is the edge length of the source image, both measured in pixels.

III. Improved ST-GCN human action recognition model with fusion attention

In modern physical education teaching, how can artificial intelligence technology be fully integrated to assist in physical education teaching and further promote the innovative development of physical education teaching content? The use of artificial intelligence technology for physical movement recognition also lays a reliable foundation for optimizing physical education teaching content and developing personalized physical education training plans. This chapter builds upon the ST-GCN model by introducing a temporal attention mechanism to construct an improved ST-GCN human movement recognition model, aiming to provide support for optimizing movement training in physical education teaching processes.

III. A. Spatio-temporal Convolutional Neural Networks and Attention Mechanisms

III. A. 1) Spatio-temporal Convolutional Neural Networks

Spatio-Temporal Graph Convolutional Neural Network (ST-GCN) extracts human skeleton from action videos by separating the two steps of spatial feature extraction and temporal feature extraction. It uses a Graph Convolutional Network (GCN) module to extract spatial features from the human skeleton graph, and then employs a Spatio-Temporal TCN module to extract temporal features of the action. In ST-GCN, the human skeleton graph signal first undergoes spatial feature extraction via the GCN module. Subsequently, the TCN module, composed of convolutional layers with a kernel size of 1×9, performs convolutional operations on the output feature maps from the GCN module. Finally, the features are fused with a residual structure to obtain the output features [21].

For human skeleton graphs, three partitioning methods are used: uniform label partitioning, distance partitioning, and spatial structure partitioning. The differences between these methods in the model lie in the different ways of decomposing the adjacency matrix. That is:

$$A + I = \sum_{i=1}^{k} A_i \tag{12}$$

In this context, k denotes the number of decomposed adjacency matrices under different partitioning schemes. The matrix $A + I$ represents the adjacency matrix of a human skeleton graph with self-connections, as described earlier, while the matrix A_i denotes the decomposed sub-adjacency matrices.

ST-GCN adopts a spatial structure partitioning method. Its GCN module implementation uses the graph convolution implementation method from existing literature. Graph convolution operations are performed on the three matrices decomposed from the adjacency matrix with self-connections, and the results are added together to obtain the GCN module's computational result.

III. A. 2) Principles of Attention Mechanisms

The concept of the attention mechanism is based on research into human vision. When humans observe certain types of objects, they often only notice key information rather than all of the object's information. Information that is not noticed is often subconsciously ignored, and this phenomenon is called the attention mechanism. Under normal circumstances, when humans receive visual signals, the brain's ability to process the information conveyed by the visual signals is not the same, and this phenomenon is called visual sensitivity [22].

The attention mechanism simulates the human brain's strategy for allocating attention. It employs a probability-based weighting approach to apply different probability weights to various regions, enabling focused attention on critical areas to achieve specific task objectives. The attention mechanism can be divided into two categories: soft attention and hard attention. Soft attention focuses on specific regions and channels within an image. This mechanism trains and identifies corresponding features through the backpropagation process of a convolutional neural network. Diverse soft attention structures can be embedded within a convolutional neural network, with their specific types depending on the network architecture design, and calculations can be performed directly based on the direction of gradients. Hard attention mechanisms, on the other hand, focus on pixels in the image as the primary concern, without using gradient directions for computation. Instead, they employ a random selection method to choose focal points. As such, the essence of attention mechanisms in selecting focal regions is to assign different weights to different positions in the input sequence, thereby prioritizing and learning from regions with higher weight values while selectively ignoring redundant information. This approach aims to allocate key information to resource processors.

III. B. Improving the ST-GCN Human Action Recognition Model

III. B. 1) Improving the ST-GCN Model Framework

To better capture changes in human body movements during sports training, this paper introduces a multi-scale spatio-temporal attention mechanism to learn spatio-temporal features of the human skeleton and seamlessly integrates them into the spatio-temporal graph convolutional network (ST-GCN) model, resulting in an improved ST-GCN model with integrated spatio-temporal attention mechanisms, enabling end-to-end training. Figure 1 shows the framework of the improved ST-GCN model. Based on the original ST-GCN module, this paper adds a multi-scale spatio-temporal attention module to extract and fuse human joint features from both temporal and spatial dimensions, thereby enhancing the global feature information of the feature maps.

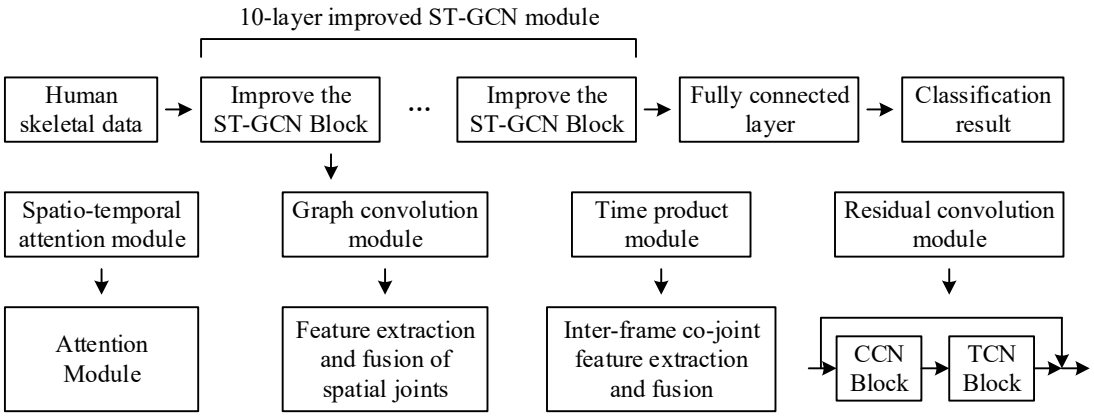


Figure 1: Overall structure of improved ST-GCN

III. B. 2) Multi-scale temporal attention

The multi-scale temporal attention module can more efficiently aggregate average pooling features and max pooling features to capture sensitive information about operations in the temporal dimension. Compared to previous multi-scale temporal modules, this module adds a shared multi-layer perceptron (S-MLP) to the module, better combining local temporal features with global temporal features. The S-MLP first shrinks and then expands features, effectively stacking multiple convolutional blocks to obtain more local temporal information, addressing the limitation of previous similar works like SENet that were insensitive to temporal features. Additionally, this module uses the Squeeze operation to compress feature dimensions, accelerating computation while reducing model computational complexity. Due to these operations, subsequent aggregation of local temporal features can be performed using simple 1D convolutions.

Given an intermediate feature $X \in R^{N\times T\times C\times H\times W}$ from the upper-level network output, use the max pooling layer to process and obtain global pooling information $F_{\max} \in R^{N\times T\times 1\times 1\times 1}$. Then, the global pooling information $F_{\max}$ is output to a shared multi-layer perceptron, which has a structure similar to that in the spatial difference module in the previous chapter, but the scaling factor is different, both being r. After the shared multi-layer perceptron processes $F_{\max}$, the intermediate spatio-temporal feature $F_{temp1} \in R^{N\times T\times 1\times 1\times 1}$ is obtained. The above operations can be summarized as:

$$F_{temp1} = \sigma(Smlp(Max(X))) \tag{13}$$

The input intermediate features X are pooled using a 3D average pooling layer to obtain the average pooling features $F_{avg} \in R^{N\times T\times 1\times 1\times 1}$. Next, the average pooling features are passed through the same shared multi-layer perceptron, where the multi-layer perceptron shares parameters with the previous path, to generate the second attention feature $F_{temp2} \in R^{N\times T\times\times 1\times 1\times 1}$, which can be represented as:

$$Avg(x) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X[:,:,:,i,j] \tag{14}$$

$$F_{temp2} = \sigma(Smlp(Avg(X))) \tag{15}$$

The obtained features $F_{\max}$ and F_{avg} are each compressed to obtain $F_{\max'}$ and $F_{avg}^{'}$, respectively, where both features have a dimension of N×T×1. Then, through a dimension concatenation operation, a new aggregated feature $F_{ios} \in R^{N\times 2T\times 1}$ is generated. Through a 1D convolution, the aggregated average pooling and max pooling features F_{ios} are further enhanced, where the convolution kernel size of the 1D convolution is 3×1. Subsequently, an Unsqueeze operation is used to restore the original feature dimensions, and through a Sigmoid activation function, the model obtains the third and most important attention feature F_{temp3}, where $F_{temp3} \in R^{N\times T\times 1\times 1\times 1}$. The above operations can be formalized as:

$$F_{avg}^{'} = Sqz(Smlp(Avg(x))) \tag{16}$$

$$F_{\max}^{'} = Sqz(Smlp(Max(x))) \tag{17}$$

$$F_{temp3} = \sigma(Unsqz(Conv([F_{avg}^{'}, F_{max}^{'}]))) \tag{18}$$

After summing the three feature matrices and then performing a dot product with the original features, the final multi-scale temporal attention features are obtained. In summary, combining the above formulas, temporal attention feature extraction can be simplified to:

$$Y = X + X\cdot(H(x) + G(x) + Z(X)) \tag{19}$$

III. B. 3) Multi-scale temporal convolutional networks

To improve the performance of TCN in both time and space, and to enhance the training efficiency and performance of deep networks, this paper incorporates residual units into the commonly used TCN and selects convolutional kernels of different time lengths. Selecting convolutional kernels of different lengths enables the proposed learning network to capture dynamic feature signals from inputs of varying durations. Meanwhile, residual connections ensure that the deep network maintains robust performance, achieving both improved performance and faster learning speeds while maintaining accuracy. The TCN with residual units performs the following operations:

$$X_l = X_{l-1} + F(W_l, X_{l-1}) \tag{20}$$

$$F(W_l, X_{l-1}) = W_l * \sigma(X_{l-1}) \tag{21}$$

Among these, the residual unit is F. For the first layer, the output of the previous layer is X, which is also the input of the first layer. The weight value is W and the nonlinear activation function is σ. Setting the ReLU function as the activation function, this function sets all negative values to zero while preserving all positive values. This allows the network parameters to be optimized through continuous forward and backward learning cycles, resulting in the following rewritten equation:

$$W * \sigma(X_{i-1}) = W * \max(0, X_{i-1}) \tag{22}$$

Therefore, when N residual units are used, the output of the hidden layer is:

$$X_N = X_1 + \sum_{i=2}^{N} W_i * \max(0, X_{l-1}) \tag{23}$$

$$X_1 = W_1 X_0 \tag{24}$$

In this case, the original input is X, which is the feature sequence formed by the joint angles of each dimension. From the above equation, it can be seen that the output X_N of the entire network is closely related to X_1. After

the second layer, as the values of X approach zero, the weight parameters of that layer are selected for discard, thereby forming residual connections to optimize the network's performance and structure.

To enable the network to learn dynamic features more efficiently across different time durations and thereby perform classification, this paper introduces improvements in the temporal network aspect. First, the initial input X undergoes a convolution operation with a duration of d_t, a total of N convolution kernels, and a stride of 1, which is then fed into the temporal convolution network. Second, the TCN is modified by selecting only its encoding layer, and the TCN with a duration of is used to process X. The results from the fully connected layer of the encoding layer are then correlated and subjected to global average pooling. Finally, the output values are classified using a SoftMax classifier to obtain the recognition results.

III. C. Validation of the effectiveness of human motion recognition models

III. C. 1) Training results of the recognition model

The experiment was conducted using a Windows 10 system, with an NVIDIA GeForce RTX 3060 (6GB) GPU and an Intel(R) Core(TM) i3-10100 CPU @ 3.60GHz 3.60 GHz processor. The experimental environment utilized the Python and PyTorch deep learning frameworks.

The improved ST-GCN human action recognition model designed in this paper was trained with 15k training iterations, an input size of 512×512, an initial learning rate of 0.001, and a batch size of 20. The curve of the loss function value during model training is shown in Figure 2.

As shown in the figure, the learning curve gradually decreases with each iteration, and the output almost entirely becomes 0. If left unchanged, it is believed that the reason is that the number of bits processed by the average loss has decreased, and the gradient has effectively disappeared. After 1.5×10^4 iterations, the training loss drops to 0.012, and the validation loss drops to 0.013. It can be seen that during the first 0.3×10^4 iterations of training on the training set, the loss value decreases rapidly due to the weights being initialized with random values. After the subsequent 1.2×10^4 iterations, the weights representing human skeleton features are fine-tuned, causing the loss value to decrease slowly and eventually stabilize. From the convergence curve, it can be concluded that the network model training has achieved the expected results.

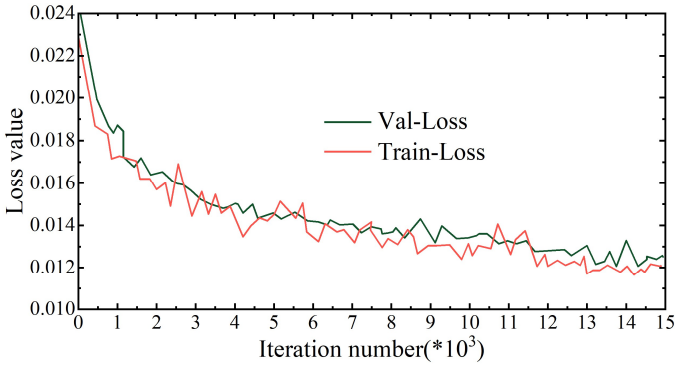


Figure 2: Model training and validation loss curves

The tests with and without scale search are denoted as maximum accuracy and 1-scale, respectively. This analysis uses the same images and batch size of 1 for each algorithm. Each analysis was repeated 500 times, and the average was taken. The performance of the improved OpenPose model, Alpha-Pose, and improved ST-GCN model in human action recognition under the same conditions is shown in Figure 3. As can be seen from the figure, the improved OpenPose model, Alpha-Pose, and improved ST-GCN, as top-down human action recognition models, have runtime that is proportional to the number of human detectors extracted. In contrast, the inference time of the improved ST-GCN model during human action recognition changes relatively little with the number of human actions. This to some extent demonstrates the efficiency of the model in this paper for human action estimation, providing support for sports training instruction.

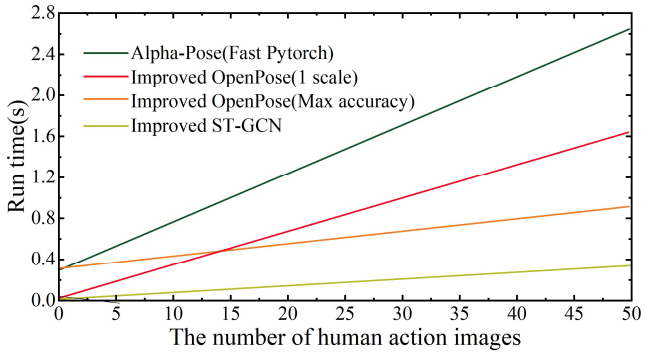


Figure 3: Comparison of model inference time

III. C. 2) Performance Comparison of Different Models

This paper collects sports competition video data through self-shot and online downloads, including single-person and multi-person scenes. Then, the OpenPose model is used to automatically label human joint points, with each complete human body consisting of 14 joint points. Finally, these preliminarily labeled data are manually proofread. The dataset constructed in this paper contains a total of 16,000 frames, of which 14,000 frames are used for training and 2,000 frames are used for testing. Additionally, the PoseTrack dataset was selected for comparison.

First, the proposed algorithm was compared with four existing advanced pose estimation algorithms—CPM, LSTM-PM, SimpleBaseline, and HRNet—on the self-constructed sports competition dataset. Table 1 shows the comparison results on the self-constructed sports competition dataset.

As shown in the table, the proposed algorithm achieved an AP of 79.3% for each keypoint and a MAP of 79.3% for all keypoints. Notably, the proposed algorithm achieved a MAP of 79.3%, demonstrating the most advanced estimation performance compared to existing methods, with improvements of 9% and 6.8% over the image-based pose estimation algorithms SimpleBaseline and HRNet, respectively. This fully demonstrates that sequence-level modeling of the collected video information and introducing additional local spatio-temporal information into the model can effectively enhance the richness and completeness of the input information. Additionally, the algorithm in this paper demonstrates significant improvements in accuracy for joints that are difficult to estimate, such as the wrist and ankle, with AP values of 80.8% and 66.8%, respectively, representing improvements of 5.4% and 10% compared to HRNet. The above results further underscore the importance of fully leveraging local sequence information and multi-scale spatial information.

Table 1: Comparison results of the self-built data set (%)

Model	CPM	LSTM-PM	SimpleBaseline	HRNet	Ours
Head	69.1	71.5	78.7	81.5	84.6
Shoulders	60.5	65.2	73.6	75.3	80.5
Elbow	68.4	69.4	79.3	81.2	85.1
Wrist	66.3	67.3	73.5	75.4	80.8
Crotch	54.6	59.8	65.2	66.8	81.4
Knee	59.3	63.4	67.8	70.2	75.9
Ankle	49.7	51.8	54.1	56.8	66.8
MAP	61.1	64.1	70.3	72.5	79.3

To further evaluate the model's performance, this paper conducted validation on the PoseTrack public benchmark dataset. Table 2 presents the accuracy rates for each keypoint and the final average accuracy rate of the proposed algorithm compared to other pose estimation algorithms, including PoseTracker, PoseFlow, FastPose, SimpleBaseline, STEmbedding, and HRNet. As shown in the table, the algorithm designed in this paper demonstrates the best estimation performance, achieving a final average accuracy of 80.5%, which is 3.4% higher than HRNet. For the challenging wrist and ankle joints, the algorithm achieves high prediction accuracy. The algorithm maintains high pose estimation accuracy in complex scenarios such as occlusion, motion blur, and changes in background lighting, and exhibits good robustness.

Table 2: Comparison results of the PoseTrack data set (%)

Model	Head	Shoulders	Elbow	Wrist	Crotch	Knee	Ankle
PoseTracker	67.4	70.4	62.4	51.5	60.8	58.5	50.5
PoseFlow	66.3	73.5	68.5	61.2	67.4	67.3	61.3
FastPose	80.2	80.6	69.3	59.3	71.2	67.9	59.5
SimpleBaseline	81.9	83.1	80.2	72.5	75.6	74.2	67.3
STEmbedding	83.7	81.7	77.5	70.8	77.9	73.9	70.2
HRNet	82.3	83.9	80.9	73.6	75.2	75.2	68.4
Ours	85.9	84.2	81.7	75.1	89.4	76.1	70.9

III. C. 3) Ablation experiment results of the model

To validate the impact of the multi-scale temporal attention (A) and multi-scale temporal convolutional network modules (B) on the model's feature extraction capability and human keypoint detection accuracy, ablation experiments were conducted to assess the effects of changes in the backbone network structure. The evaluation metrics selected were the percentage of correct keypoints (PCK), the number of parameters, and the average precision rate. PCK is defined as the proportion of correctly estimated keypoints, calculated as the ratio of detected keypoints whose normalized distance to their corresponding labels is below a set threshold. Table 3 presents the ablation experiment results for the model.

According to the experimental results in the table, the baseline model without the multi-scale temporal attention mechanism and multi-scale spatio-temporal convolutional network has a parameter count of 21.4 MB. Adding the multi-scale temporal attention mechanism and multi-scale spatio-temporal convolutional network increased the parameter counts by 6.07% and 5.14%, respectively. The overall parameter count of the improved ST-GCN model increased by 8.41%. The addition of the attention mechanism and spatio-temporal convolutional network significantly improved mPCK and MAP. Although this increases the model's parameter count to some extent, the improved accuracy results justify the increase. Introducing the attention mechanism does not significantly impact the model's computational complexity, and adding the attention mechanism can moderately improve the model's accuracy in detecting keypoints.

Table 3: The ablation experiment results of the model

Model	Parameter/MB	mPCK	MAP/%
Baseline	21.4	0.843	78.5
Baseline+A	22.7	0.909	85.3
Baseline+B	22.5	0.914	86.1
Ours	23.2	0.932	88.2

IV. Sports motion evaluation analysis combining the DTW algorithm

In the field of education, the rapid development of artificial intelligence can drive systematic change in the education sector and promote internal renewal and transformation of the education system. The introduction of artificial intelligence technology can help students quickly master basic sports knowledge and skills in physical education, and will also make sports training more scientific and efficient, contributing to the visualization and personalization of physical education.

IV. A. Action evaluation based on the DTW algorithm

IV. A. 1) Dynamic Time Warping Algorithm

A time series is an ordered set of feature values observed at fixed time intervals in sequential order. Feature values can be one-dimensional or multi-dimensional. Let $X=\{x_1,x_2,\cdots,x_n\}$ and $Y=\{y_1,y_2,\cdots,y_m\}$ are two one-dimensional time series of lengths n and m, respectively. The distance between points x_i and y_j is defined as $d(i,j)=|x_i-y_j|^p$, where $||^p$ denotes the pth norm. Create an $n\times m$ distance matrix, where the element (i,j) represents the alignment of the ith point in X with the jth point in Y. Find a path through several grid points in the matrix, which represents the matching relationship between each point in sequences X and Y. Different paths correspond to different matching relationships.

The DTW algorithm finds a curved path such that the sum of distances between all matched point pairs along the path is minimized. This minimum distance sum corresponds to the DTW distance, and the path is called the optimal curved path. Let the curved path be R, where R is an ordered set containing K binary arrays, i.e.:

$$R = \{(i_1, j_1), (i_2, j_2), \cdots, (i_k, j_k), \cdots, (i_K, j_K)\} \tag{25}$$

The curved path R must satisfy the following three conditions:

(1) The boundary conditions are $i_1 = j_1 = 1, i_K = n, j_K = m$.

(2) Monotonicity: $i_k \le i_{k+1}, j_k \le j_{k+1}$.

(3) Continuity: $i_k + 1 - i_k \le 1, j_{k+1} - j_k \le 1$.

The formula for calculating the DTW distance is:

$$DTW_p(X,Y) = \sqrt[p]{\min\left\{\sum_1^K d(i_k, j_k)\right\}} \tag{26}$$

Based on the constraints, use dynamic programming to construct a cumulative distance matrix D. The above problem can then be converted into the following solution:

$$DTW_p(X,Y) = \sqrt[p]{D(m,n)} \tag{27}$$

$$D(i,j) = d(i,j) + \min\begin{cases} D(i-1,j) \\ D(i,j-1) \\ D(i-1,j-1) \end{cases} \tag{28}$$

Among these, $D(0,0) = 0, D(i,0) = D(0,j) = \infty, i = 1, \cdots, n, j = 1, \cdots, m$, The cumulative distance $D(i,j)$ is the sum of the current element's distance $d(i,j)$ and the cumulative distance to the smallest neighboring element that can be reached from it. The final cumulative distance $D(m,n)$ corresponds to the DTW distance between sequences X and Y, which is a measure of the similarity between time series X and Y. For two time series of lengths n and m, the time complexity of the DTW distance is $O(m \times n)$.

IV. A. 2) Sports motion evaluation based on DTW

There are significant differences in the difficulty of various movements in physical education instruction. When evaluating movements, specific evaluation methods must be applied to specific types of movements. During physical movement evaluation, the time taken by most performers to complete a movement typically deviates from the standard time, necessitating the alignment of recorded movements. The DTW algorithm enables temporal alignment and matching of movements, facilitating accurate analysis and comparison. This study investigates the application of the DTW algorithm for movement alignment in physical training data.

The research design for physical education training movement evaluation includes three sub-items: movement pre-evaluation, movement segmentation, and movement detail evaluation. Movement pre-evaluation can detect issues such as missing movements and disordered movement sequences in movement sequences. First, the Euclidean distance vector between the standard posture and the user's movement sequence posture is calculated. That is:

$$d_i = \| b_1 - p_i \|_2 \tag{29}$$

In the equation, d_i represents the Euclidean distance between the i th pose and the first standard pose in the standard action library, b_1 represents the joint angle vector of the first standard pose, and p_i represents the joint angle vector of the i th pose of the user.

Action segmentation involves generating an Euclidean curve using the Euclidean distance vector, adding a positive offset to the minimum Euclidean distance, calculating the corresponding start frame and end frame, and using the start frame and end frame for action segmentation. Action detail evaluation assesses action quality based on joint angles, action center time, action duration, and dynamic joint average angular velocity. The similarity of static action joint angles can be expressed as:

$$d_s = \| \frac{1}{n}(c_1 + c_2 + \cdots + c_n) - b \| \tag{30}$$

In the formula, d_s represents the joint similarity of static movements, n represents the total number of frames of static movements, c represents the joint angle vector in static movements, and b' represents the standard posture vector. When calculating the joint angle similarity of dynamic movements, first calculate the Euclidean distance between the user's posture and the standard movement posture, that is:

$$d_{ij} = \| e_i - e_j^{'} \|_2 \tag{31}$$

In the formula, d_{ij} represents the Euclidean distance between the i th pose of the user and the j th pose of the standard action, and e represents the joint angle vector. The cumulative matrix is obtained using the distance matrix, and then the shortest path is calculated, i.e.:

$$g_{ij} = d_{ij} + \min\{g_{i-1}, g_{i,j-1}, g_{i-1,j-1}\} \tag{32}$$

In the formula, g_i represents the shortest path from the first row and first column of the accumulation matrix to the i th row and j th column. The similarity of dynamic joint angles is calculated as follows:

$$d_d = \frac{g_{nm}}{n+m} \tag{33}$$

In the formula, d_d represents the similarity of dynamic motion joint angles. When calculating the similarity of motion center times, first calculate the similarity of the user's motion center times, that is:

$$\begin{cases} t_c = \frac{1}{2}(f_{start} + f_{end}) \times \frac{1}{\tau} \\ e_c = t_c - t_c^{'} \end{cases} \tag{34}$$

In the formula, t_c represents the center time of the user's action, f_{smart} represents the start frame of the action, f_{end} represents the end frame of the action, τ is the sampling frequency of the device, and $t_c^{'}$ represents the center time of the standard action. The expression for calculating the similarity of action duration is:

$$\begin{cases} t_s = (f_{start} + f_{end}) \times \frac{1}{\tau} \\ e_s = t_s - t_s \end{cases} \tag{35}$$

In the equation, t_s and $t_s^{'}$ represent the duration of the user's action and the standard action, respectively, and e_s represents the similarity of the action duration. The similarity of the average angular velocity of the joints is calculated as follows:

$$e_w = \frac{1}{n-1} \| \tau(e_{i+1} - e_i) \|_1 - w \tag{36}$$

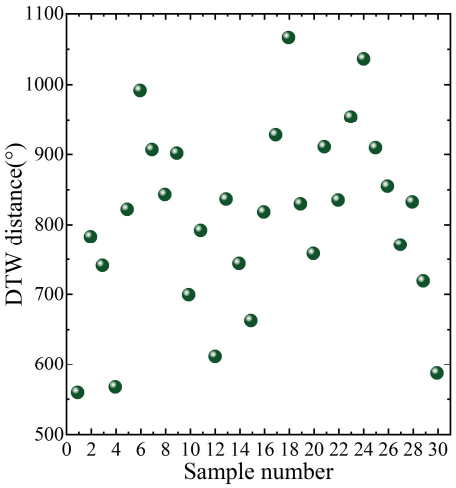
In the equation, e_i represents the pose joint angle vector, $w^{'}$ represents the standard angular velocity, and n represents the total number of frames in the motion sequence.

IV. B. Analysis of sports movement evaluation results

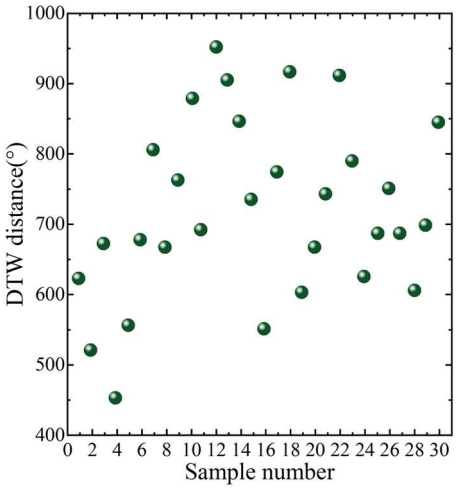
IV. B. 1) Joint Angle Distance in Sports Movements

Motion evaluation aims to compare test motions with standard motions to assess the performance of the test motions. Traditional motion evaluation methods involve manual observation and assessment by experienced personnel such as coaches and referees; however, the results of such evaluations may be influenced by subjective factors. To achieve fair and impartial scoring, computer-assisted scoring can be considered, utilizing the principle of similarity. The DTW algorithm is employed to calculate the distance between the test motion and the standard motion sequence across eight joint angles, which serves as the similarity assessment parameter. The front punch from the dataset of ten actions is selected as the evaluation object, with 30 samples chosen as test sequences. Action video data from sports majors is used as the standard action sequence template. Figures 4–7 show the DTW distance distributions for the left shoulder, left elbow, right shoulder, right elbow, left knee, left hip, right knee, and right hip, respectively.

As shown in the figures, the DTW distances for the left shoulder joint angle are mostly distributed between 550° and 1000°, with a small portion between 1000° and 1100°, which is discarded. The DTW distances for the left elbow joint angle are mostly distributed between 550° and 1000°, with a small portion between 400° and 550°, which is discarded. The DTW distances for the right shoulder joint angle are distributed between 120° and 280°, while those for the right elbow joint angle are distributed between 135° and 280°. The DTW distance distributions for the left knee and left hip joint angles, as well as the right knee and right hip joint angles, can be obtained from the figure. From the DTW distances in the figure, it can be seen that the four joint angles of the upper limbs have a relatively sparse distribution, while the four joint angles of the lower limbs have a relatively dense distribution, indicating that the range of motion of the upper limbs is larger.

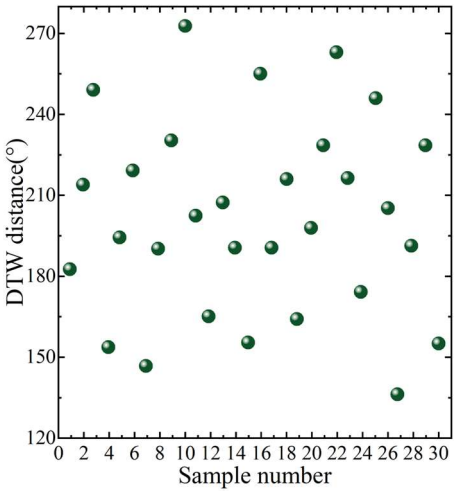


(a) Left shoulder

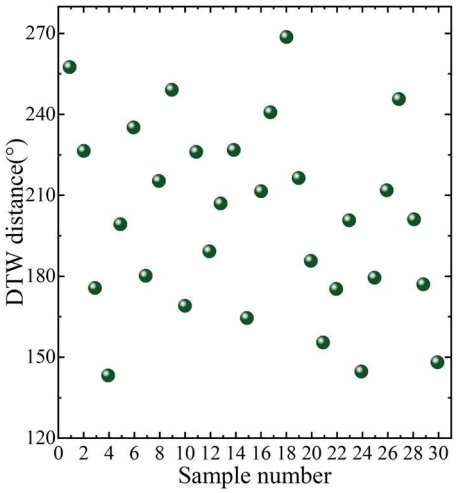


(b) Left elbow

Figure 4: DTW distance distribution of left shoulder and left elbow



(a) Right shoulder



(b) Right elbow

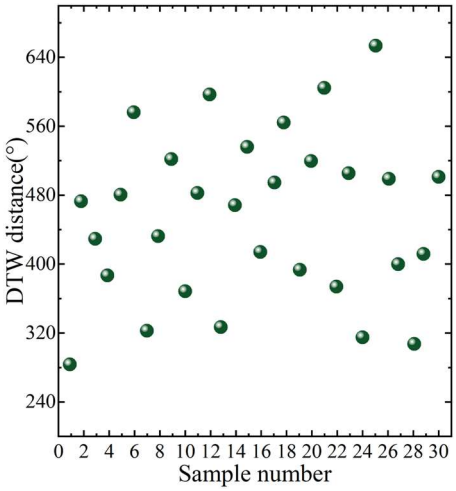
Figure 5: DTW distance distribution of right shoulder and right elbow

IV. B. 2) Sports Movement Evaluation Results

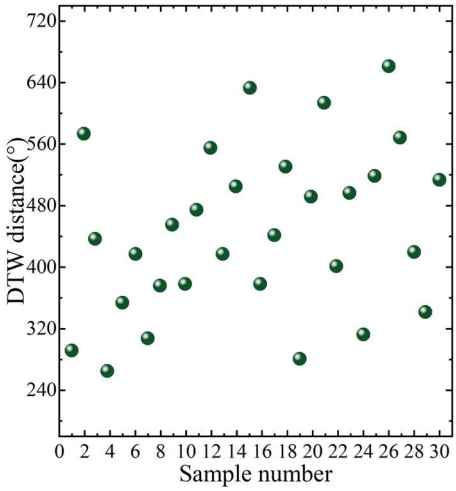
Based on the above experimental analysis, an evaluation method was constructed for sports movements, and the movement score was calculated using $S_a = S_c - (d_1 - d_2) \times f_c$ In the equation, S_a represents the score for an angle feature, S_c represents the score for angle allocation. d_1 and d_2 represent the DTW distance value and the minimum value within the effective interval of the DTW distance, respectively. f_c is the loss parameter, whose value is related to the amplitude of the action change. Joints with larger amplitudes have smaller loss parameters, while those with smaller amplitudes have larger loss parameters. The final total score is the sum of the scores for the eight joint angles: $S = \sum_{a=1}^{8} S_a$.

Each joint angle has a DTW distance distribution interval. After multiple experiments, the baseline DTW distance values and joint loss parameters for different joint angles were obtained, and these were then substituted into the formula to calculate the action evaluation score. The action evaluation results are shown in Table 4. As shown in the table, the total evaluation score obtained using the method proposed in this paper is 88.19 points, which is only 1.12 points different from the professional scoring results. This indicates that using the DTW distance based on the

joint angle curve of sports movements as a parameter, the action evaluation method has been validated through action assessment, demonstrating its rationality.

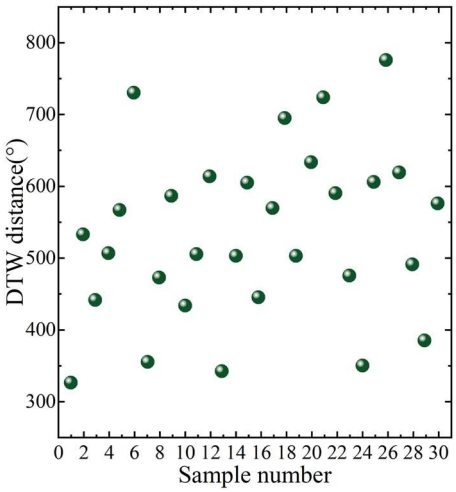


(a) Left knee

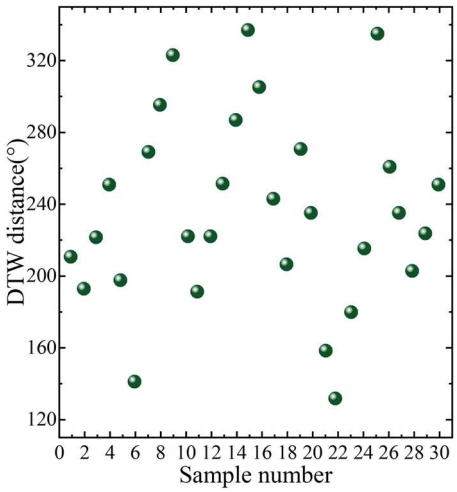


(b) Left crotch

Figure 6: DTW distance distribution of left knee and left crotch



(a) Right knee



(b) Right crotch

Figure 7: DTW distance distribution of right knee and right crotch

Table 4: Action evaluation results

Joint Angle	DTW distance	Evaluation score	Professional score
Right shoulder	963	10.42	10.18
Left shoulder	1005	10.98	10.54
Right elbow	1123	11.63	12.26
Left elbow	1294	10.76	11.37
Right crotch	463	11.53	11.94
Left crotch	472	11.42	10.98
Right knee	569	11.17	11.35
Left knee	738	10.28	10.69

Relying on sports movement scoring methods can help physical education teachers grasp the standardization of students' movements during sports activities. Combined with the human movement recognition model mentioned above, it can ensure the richness of physical education teaching content and provide a reform path for personalized training in physical education teaching.

V. Conclusion

The paper uses OpenPose to obtain skeletal point data for sports movements, then introduces a multi-scale temporal attention and residual-enhanced ST-GCN model to construct a human motion recognition model, and combines it with the DTW algorithm for sports movement evaluation analysis. The results show that the MAP value of the improved ST-GCN model is 79.3%, which is 9% and 6.8% higher than the image-based pose estimation algorithms SimpleBaseline and HRNet, respectively. Sports movements evaluated using the DTW algorithm scored 88.19 points, differing by only 1.12 points from manual scoring. Therefore, integrating artificial intelligence technology with sports education and teaching can promote the intelligent development of sports education and teaching, and also provide support for personalized training in sports education.

Funding

This work was supported by Major Project of Scientific Research Plan Compilation of Anhui Province (Project Number: 2024AH04032).

References

[1] Darling-Hammond, L. (2009). Teaching and educational transformation. Second international handbook of educational change, 505-520.
[2] Østergaard, L. D. (2016). Inquiry-based learning approach in physical education: Stimulating and engaging students in physical and cognitive learning. Journal of Physical Education, Recreation & Dance, 87(2), 7-14.
[3] Simonton, K. L., Layne, T. E., & Irwin, C. C. (2021). Project-based learning and its potential in physical education: an instructional model inquiry. Curriculum Studies in Health and Physical Education, 12(1), 36-52.
[4] Adolf, V. A., Kondratyuk, A. I., Kondratyuk, T. A., Sitnichuk, S. S., Zaitseva, M. S., & Kolesova, N. V. (2021). Professional training of Physical Education teachers in digital transformation of society. European Proceedings of Social and Behavioural Sciences.
[5] Castillo, I., Molina-García, J., Estevan, I., Queralt, A., & Álvarez, O. (2020). Transformational teaching in physical education and students' leisure-time physical activity: The mediating role of learning climate, passion and self-determined motivation. International journal of environmental research and public health, 17(13), 4844.
[6] Gholami, A. (2024). The Integration of Spiritual Curriculum into Primary School Physical Education in Relation to Transformational Teaching Approaches. Physical Activity in Children, 1(1), 14-20.
[7] Ward, P. (2013). The role of content knowledge in conceptions of teaching effectiveness in physical education. Research Quarterly for exercise and sport, 84(4), 431-440.
[8] Zhang, Y. K. (2017). Transformation research on teaching practice mode of physical education major based on applied-oriented talents cultivation. Eurasia Journal of Mathematics, Science and Technology Education, 13(10), 7089-7097.
[9] Sun, J. (2021). Design and Application of Intelligent Teaching System of College Physical Education Based on Artificial Intelligence. In Frontier Computing: Proceedings of FC 2020 (pp. 1729-1736). Springer Singapore.
[10] Song, X. (2024). Physical education teaching mode assisted by artificial intelligence assistant under the guidance of high-order complex network. Scientific Reports, 14(1), 4104.
[11] Yang, D., Oh, E. S., & Wang, Y. (2020). Hybrid physical education teaching and curriculum design based on a voice interactive artificial intelligence educational robot. Sustainability, 12(19), 8000.
[12] Lee, H. S., & Lee, J. (2021). Applying artificial intelligence in physical education and future perspectives. Sustainability, 13(1), 351.
[13] Wang, C., & Wang, D. (2023). Managing the integration of teaching resources for college physical education using intelligent edge-cloud computing. Journal of Cloud Computing, 12(1), 82.
[14] Hu, Z., Liu, Z., & Su, Y. (2024). AI-Driven Smart Transformation in Physical Education: Current Trends and Future Research Directions. Applied Sciences, 14(22), 10616.
[15] Li, S., Wang, C., & Wang, Y. (2024). Fuzzy evaluation model for physical education teaching methods in colleges and universities using artificial intelligence. Scientific Reports, 14(1), 4788.
[16] Liu, T., Wilczyńska, D., Lipowski, M., & Zhao, Z. (2021). Optimization of a sports activity development model using artificial intelligence under new curriculum reform. International Journal of Environmental Research and Public Health, 18(17), 9049.
[17] Liu, G. (2022). Physical education resource information management system based on big data artificial intelligence. Mobile information systems, 2022(1), 3719870.
[18] Ding, D., Shen, Y., Jiang, J., Yuan, Q., Xiu, T., Ni, K., & Liu, C. (2023). Data collection and information security analysis in sports teaching system based on intelligent sensor. Measurement: Sensors, 28, 100854.
[19] Kailai Jiang. (2025). Optimization study of badminton sports training system based on MoileNet OpenPose lightweight human posture estimation model. Entertainment Computing,54,100975-100975.
[20] Baojun Liu & Zimin Wang. (2024). Research on human behaviour recognition method of sports images based on machine learning. International Journal of Bio-Inspired Computation,23(2),99-110.
[21] Qi Lu. (2024). Sports-ACtrans Net: research on multimodal robotic sports action recognition driven via ST-GCN. Frontiers in Neurorobotics,18,1443432-1443432.
[22] Jinmao Tong & Fei Wang. (2024). Basketball Sports Posture Recognition Technology Based on Improved Graph Convolutional Neural Network:Regular Papers. Journal of Advanced Computational Intelligence and Intelligent Informatics,28(3),552-561.