International Journal of Innovative Computing, Information and Control
Volume **21**, Number **1**, February **2025**

 ISSN 1349-4198
pp. **37–52**

DESIGN OF SPORTS ACTION RECOGNITION AND EVALUATION BASED ON IMPROVED DTW ALGORITHM

Yuli Hu[1,*] and Di Liu[2]

[1]Aviation Fundamentals College
[2]School of Engineering
Naval Aviation University
Yantai 264001, P. R. China
*Corresponding author: YuliHu56@163.com

Received March 2024; revised July 2024

Abstract. *With the popularization of sports and the development of computer technology, the demand for sports action recognition and evaluation is increasing day by day. Although existing methods have achieved certain results, there are still shortcomings in recognition accuracy, real-time performance, and stability. To improve the effectiveness of sports action recognition and evaluation, this study proposes an action recognition and evaluation method based on an improved dynamic time warping algorithm. It utilizes an improved 3D convolutional network (C3D-Resnet) to extract sports action features, and combines feature fusion and dimensionality reduction methods to improve the dynamic time warping algorithm. The test results on the CASIA TaiChi Dataset showed that the accuracy of C3D, Resnet, and C3D-Resnet were 84.5%, 86.4%, and 94.7%, respectively. After feature data dimensionality reduction, the data dimension decreased from [289, 678] to within the range of [17, 109], and the average action recognition rate increased from 88.3% to 91.2%. The F1 value of the improved dynamic time regularization algorithm was about 97.1%, and the difference between the sports action evaluation results and the scores of professional coaches was less than 1 point. This study has achieved accurate recognition and evaluation of sports movements, which has important application value and practical significance in improving the effectiveness of sports training and teaching.*
Keywords: Sports movements, Feature extraction, Identification and evaluation, KNN, DTW

1. **Introduction.** Sports occupy a meaningful position in modern society, not only as a form of entertainment, but also as a way to promote health and develop personal skills [1]. The continuous development of technology has deepened people's understanding of sports. The Identification and Evaluation of Sports Movements (IEoSM) is a hot research topic. It can not only provide objective evaluation of athlete performance, but also provide strong guidance for training and skill improvement [2]. However, there are still some challenges in identifying and evaluating sports movements at present. The traditional approach often relies on manual observation and subjective evaluation, which has problems such as strong subjectivity and inconsistent judgment standards [3]. For complex sports movements, such as transient and continuous movements, traditional methods are often difficult to accurately capture and analyze [4]. Among the existing sports action recognition algorithms, the Dynamic Time Warping (DTW) algorithm is widely used for matching and aligning time series data, with good performance. However, traditional DTW algorithms have certain limitations in terms of computational complexity and recognition accuracy. Video sequences are stacked by a series of Red, Green and Blue (RGB) images along the temporal

DOI: 10.24507/ijicic.21.01.37

dimension. It contains large variations and complexities compared with images, so it is difficult to model an effective spatio-temporal representation [5]. Therefore, this study proposes an IEoSM design based on an improved DTW algorithm. Compared with the traditional DTW, the improved algorithm has innovation in two aspects. Firstly, this study introduces the K-Nearest Neighbor (KNN) algorithm to enhance recognition accuracy by classifying the nearest neighbors of samples. Secondly, by optimizing the DTW algorithm, the computational complexity has been reduced and the algorithm's real-time performance has been lifted. The innovation of this study is mainly reflected in the combination of DTW and KNN algorithms to obtain KNN-DTW, which fully utilizes the advantages of both algorithms and lifts the stability of the algorithm. This study aims to provide an innovative method for the IEoSM field and technical support for athlete training. The following content is divided into four parts. Section 2 is a summary of relevant research. Section 3 is about designing the IEoSM method and it is validated in Section 4. The last section is a summary of the entire study.

2. **Related Works.** Human Motion Characteristics (HMC) refer to the characteristic information that can be used to describe and recognize actions, such as posture, Action Sequence (AS), and joint angle, exhibited by the human body during various activities. These features can be extracted through methods such as motion capture, image processing, and deep learning for applications such as human motion recognition, analysis, and evaluation. Zhao et al. combined static image feature extraction with deep learning to process video data, and extracted intra frame feature vectors through pre trained deep networks. This method was superior to traditional Long Short-Term Memory (LSTM) models [6]. Ren et al. proposed a single dual stream ConvNets framework that utilizes multimodal learning of RGB and deep stream feature extraction to represent RGB-D sequences as motion images as inputs to the proposed ConvNets for obtaining spatio-temporal information. The effectiveness of this framework reached 90% [7]. Chen's team proposed a video based action recognition network that combines the advantages of 2D Convolutional Neural Network (CNN) and 3D CNN to learn temporal motion features. The accuracy could reach 93.1% [8]. Xue's research team proposed a self supervised learning method built on mutual information to learn action features from videos without manual annotation. For action recognition tasks in videos, this method could serve as an effective pre training tool [9]. Gao's team proposed a multi-dimensional data model for video image motion recognition and captured based on a deep learning framework. It used a Gaussian mixture model to extract the motion foreground of the target and combined it with gradient histograms to recognize the human body, with an average classification accuracy of 85.79% [10].

IEoSM refers to the use of posture recognition and motion analysis on the human body to determine and evaluate the correctness, accuracy, and quality of sports movements, providing scientific basis for sports training and competition. Yu et al. applied computer video processing technology to evaluating the training efficiency of table tennis players, and combined it with video processing to identify. This method could improve athlete training efficiency by 25% [11]. Nguyen's team members proposed a dual feature dual motion network for sports action skeleton recognition, and enriched spatial information through branch networks. The recognition time of this method was 0.04 milliseconds [12]. Zhao et al. proposed a wearable sensor with multiple inertial measurement units, which establishes a mapping relationship between emotions and human motion through fuzzy comprehensive evaluation, and recognizes human motion emotions. The accuracy of this method could reach 90% [13]. Xu et al. proposed an attention based gait recognition network, which solves the problem of feature dilution in gait recognition by designing

new local gait representation methods and an attention model based on prior information. This method performed well on the CASIA-B and OU-MVLP datasets [14]. Wu's team proposed a potential label mining strategy for identifying group activities in basketball videos. It used unsupervised hierarchical clustering technology to extract potential labels from motion patterns and trained deep CNN for frame level feature extraction, with an accuracy rate of up to 89% [15].

In summary, many researchers have conducted different designs and studies on the extraction and evaluation of sports action features. However, there are still some shortcomings and gaps in the relevant research work. For example, the recognition accuracy of existing methods in dealing with complex scenes and diverse poses still needs to be improved, and most methods rely on deep learning, lacking effective adaptability and universality. In addition, research on sports action recognition and evaluation is not yet sufficient, especially in terms of action quality evaluation and training efficiency improvement in practical application scenarios. Therefore, the study proposes a sports action recognition and evaluation method based on the improved DTW algorithm. By combining the DTW algorithm and improved deep learning technology, the aim is to improve the accuracy, adaptability, and universality of sports action recognition and evaluation, and provide more scientific basis for sports training and competition in practical application scenarios.

3. Design of IEoSM Method Based on Improved DTW.

This chapter designs a method for extracting sports action features, including human pose detection, Gaussian distribution description of joint position, true joint position, and heat map generation. Meanwhile, an LSTM network model based on attention mechanism (Attention-LSTM) was proposed for accurate localization of sports action features. In addition, the KNN algorithm was also used for IEoSM, and the DTW algorithm was added to KNN to solve the problem of unequal time series length. An action scoring method based on bone data was proposed.

3.1. Design of sports action feature extraction method.

Sports Action Feature Extraction (SAFE) refers to extracting motion related features from sports videos or images. These features can be used for applications such as action recognition, action analysis, and action evaluation [16]. When performing SAFE, it is usually necessary to first perform human pose detection to obtain the position information of human joints, and then extract features related to specific actions by analyzing this position information [17]. To predict the position of joint points, this study uses Gaussian distribution to describe the annotated key points in the sports action dataset. The Gaussian distribution of joint points is Equation (1).

$$S^*_{j,k}(p) = \exp\left(-\frac{\|p - x_{j,k}\|_2^2}{\sigma^2}\right) \tag{1}$$

In Equation (1), the Gaussian distribution of the $S^*_{j,k}(p)$ represents joint point j at point p, and the positions of the joint points j belonging to person k marked are $x_{j,k}$, with a standard deviation of σ. The vector dimension of joint j is 3. The true position of the joint points is Equation (2).

$$S^*_j(p) = \max S^*_{j,k}(p) \tag{2}$$

In Equation (2), the true position of the joint j at point p is $S^*_j(p)$. The expression of pixels on limbs in the image area is Equation (3).

$$L^*_{c,k}(p) = v \tag{3}$$

In Equation (3), the unit vector of pixel p on the limb connection of person k is $L^*_{c,k}(p)$, and the horizontal joint component is v. The key point position in the image is c. The annotated positions of human joints are shown in Equation (4).

$$L^*_c(p) = \frac{1}{n_c(p)} \sum_k L^*_{c,k}(p) \tag{4}$$

In Equation (4), the position of pixel p in the annotation belongs to the Bth person's j joint points, which is $L^*_c(p)$, and the number of vectors at point p is n_c. By labeling the training samples for human pose estimation, a vector field map of the same size as the original image can be generated [18]. This marking method can preserve the rotation information of limbs and mark the information at occluded areas. In summary, this labeling method can obtain complete samples for human pose estimation, which can be used to train neural networks. The stage structure of the neural network is shown in Figure 1.

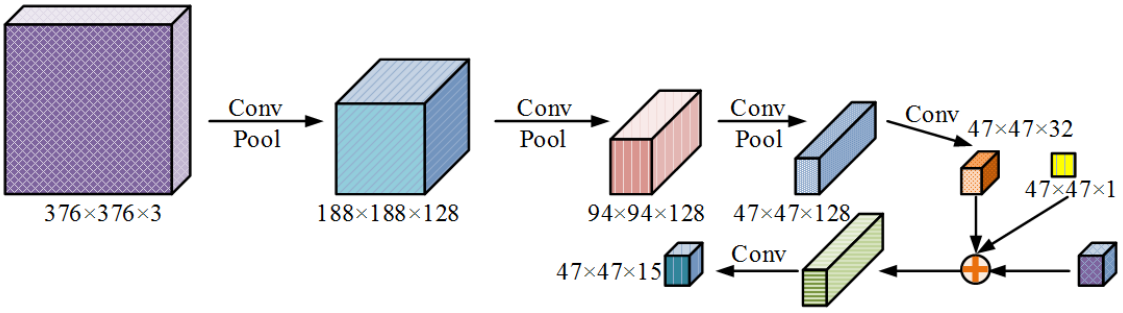


FIGURE 1. Stage structure of neural networks

The input of this neural network is the feature map F processed by the skeleton network, which has a size of $376 \times 376 \times 3$ and is extracted by the GhostNet skeleton network [19]. The output is 19 heatmaps, of which 18 are heatmaps of 18 joint points, and the other 1 is a vector field map of limb connections. The entire neural network is divided into t stages, with the same structure used in all stages except for stage 1. In the intermediate process of the neural network, F first undergoes a 1/8 downsampling operation to obtain a $47 \times 47 \times 15$ feature map. This feature map is input into the next stage along with the output of each stage. Each stage will output a $47 \times 47 \times 15$ heatmap, including 18 heatmaps of related nodes and 1 graph of limb connection vector field. The feature map of F after downsampling will be compared with the output of each stage to obtain the loss of that stage. The structure of the neural network stage includes a skeleton network and a joint prediction network. The skeleton network uses GhostNet to extract feature maps as input. The joint prediction network adopts a residual network (Resnet) structure, which outputs heat maps and limb connection vector field maps through multiple stages, and calculates losses. The neural network generates losses at each stage, and the feature point prediction loss calculation is Equation (5).

$$f^l_S = \sum_{j=1}^{J} \sum_p W(p) \left\| S^t_j(p) - S^*_j(p) \right\|^2_2 \tag{5}$$

In Equation (5), the cost function for feature point prediction is f^l_S, the predicted position of joint point j at time t is $S^t_j(p)$, the number of neural network stages is J, and the weight function is W. The cost function for limb prediction is Equation (6).

$$f_L^l = \sum_{c=1}^{C} \sum_{p} W(p) \left\| L_c^t(p) - L_c^*(p) \right\|_2^2 \tag{6}$$

In Equation (6), the cost function for limb prediction is f_L^l, and the predicted joint position of pixel p at time t is $L_c^t(p)$. The overall cost function is Equation (7).

$$f_{total} = f_S^l + f_L^l \tag{7}$$

In Equation (7), the overall cost function is f_{total}, which is taken to calculate the overall loss of training. This study uses L2 type cost functions for parameter updates. The parameter matrix and regularization parameters are taken into account. In the revised gradient backpropagation formula, the learning rate will gradually decrease with the rise of training rounds to ensure that the weights of the neural network converge to the ideal value. The environmental action features are implemented through an improved 3D convolutional network (C3D Resnet), with the main improvement being the use of Resnet to prevent gradient vanishing. The heat map predicted by neural networks in this study can identify regions with possible feature points. Using non maximum suppression methods, regions with Gaussian distribution values greater than the threshold are used as confidence regions, and the point with the highest confidence value is obtained as the joint point. Then, by determining the integration value of the unit vector and coordinate axis direction of the limb parts, the connection method of the feature points is determined, and the most accurate connection method of the feature points is obtained. This study proposes an Attention-LSTM model for SAFE. This model achieves accurate localization of HMC by combining attention mechanism and time-domain feature extraction. This method utilizes pose detection and object segmentation algorithms to extract human pose features, and uses a pre trained LSTM network for feature extraction and time localization. Through attention models, networks can more accurately extract the features of joints and weight the importance of different frames and joints, thereby improving the accuracy of action recognition. The method for determining feature points mainly includes the following steps. First, by using a modified gradient backpropagation formula and considering the parameter matrix and regularization parameters, the learning rate of the neural network will gradually decrease with the increase of training rounds to ensure that the weights converge to the ideal value. Secondly, the extraction of environmental action features is achieved through the use of improved C3D Resnet. Then, using non maximum suppression methods, the regions with Gaussian distribution values greater than the threshold are used as confidence regions, and the point with the highest confidence value is obtained as the joint point. Finally, by determining the integration value of the unit vector and coordinate axis direction of the limb parts, the connection method of the feature points is determined, and the most accurate connection method of the feature points is obtained. The LSTM structure that integrates spatio-temporal attention mechanism (SAM) is Figure 2.

In the LSTM based SAM method, the input of time period 0 is encoded with a single-layer LSTM and combined with the input of time period 1. The result of this process is processed by the tanh activation function and then input into the fully connected layer. The fully connected layer outputs a score matrix that represents the degree of influence of different frames on actions. The score output by the attention model takes effect before the output layer of the neural network, which means that the higher the score, the greater the impact of the current frame on the action classification results. In addition, spatial attention models also play a role in the input layer of neural networks. Due to the fact that the input layer receives the coordinates of the joint points, each joint point corresponds to a score. The attention score matrix is calculated as shown in Equation (8).

$$s_t = U_s \tanh\left(W_{xs} + W_{hs} h^*_{t-1} + b_s\right) + b_{us} \tag{8}$$

In Equation (8), the score matrix is s_t, the weight matrix is U_s, the parameter matrix are W_{xs} and W_{hs}, the output value of the hidden layer is h^*_{t-1}, and the deviation are b_s and b_{us}. In sports action recognition, it is very important to comprehensively consider the Environmental Action Characteristics (EAC) and HMC [20]. EAC can provide information about sports venues, equipment, and their trajectories, which can help people understand the entire scene and background, thereby better analyzing the movements of athletes. HMC, on the other hand, focuses on the athlete's own movements and postures. By analyzing and extracting the athlete's movement trajectory, joint angles, etc., it can more accurately identify and understand the athlete's movements.

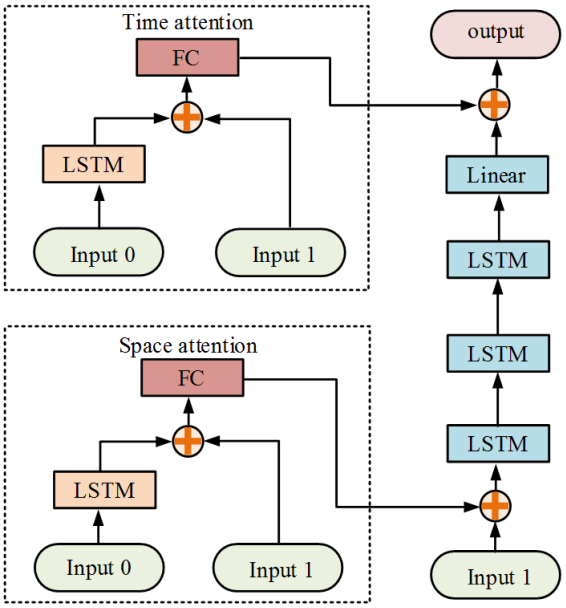
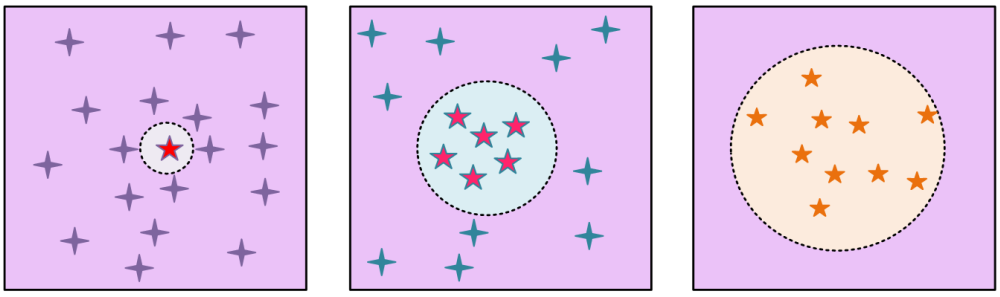


FIGURE 2. LSTM structure integrating SAM

3.2. **Design of IEoSM method.** KNN is a basic classification and regression algorithm that calculates the distance between the sample to be classified and the known sample in the training set, and then determines the category of the sample to be classified based on the size of the distance [21]. The IEoSM method can use the KNN algorithm for classification and evaluation, determining the category of samples or evaluating the degree of difference between samples and standard samples by calculating the distance between samples. The classification performance of KNN algorithm is Figure 3.



(a) Overfitting (b) Suitable fitting (c) Under-fitting

FIGURE 3. KNN algorithm classification performance

When selecting the K value, it is necessary to consider the sensitivity and generalization ability of the classifier. If the K value is too small, the classifier will be affected by errors or abnormal data, resulting in classification errors. If the K value is too large, it will vote over a larger range, causing the classifier's prediction performance to deteriorate. Therefore, the values are usually acquired using cross validation methods. Cross validation divides all ASs into K disjoint subsets, and then takes one of the K disjoint subsets as the test AS, with the remaining K − 1 as the training set for the AS. Train the model K times utilizing the training set, and the classifier's prediction results are obtained. The AS test set is tested on the trained model and the recognition rate is obtained. Finally, the final recognition rate of the classification model is the mean of the recognition rate gained from K-testing. However, the KNN algorithm has some shortcomings. Firstly, Euclidean Distance (ED) can only measure two time series of equal length, and it is hard to measure ED when two time series are not of equal length. Therefore, this study introduces the DTW algorithm distance as a distance metric in KNN to solve the problem of unequal time series, and names this method KNN-DTW. In addition, dimensionality reduction of data samples can also reduce computational complexity. The KNN-DTW regular path is shown in Figure 4.

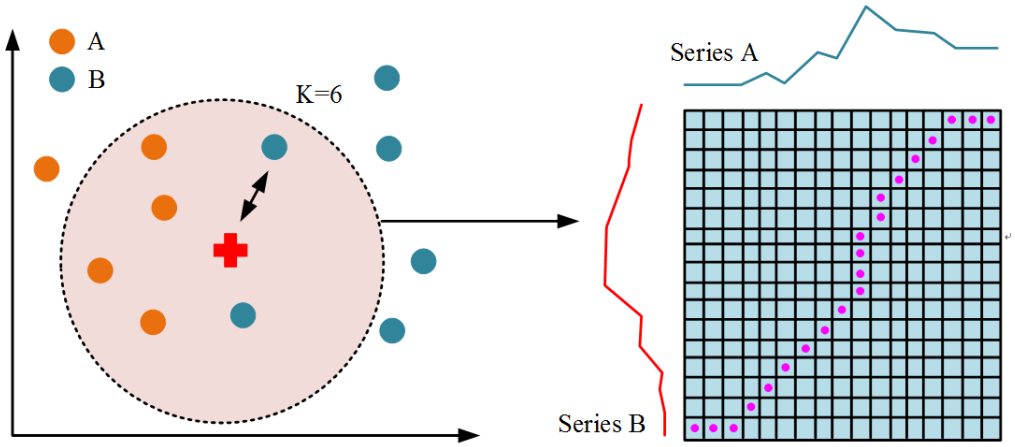


FIGURE 4. KNN-DTW regularization path

In Figure 4, sequences A and B represent two time series that exhibit significant similarity after scaling transformation. DTW can make two ASs as long as possible and calculate the cumulative distance between them. The representation of sequence A is Equation (9).

$$A = (A_1, A_2, \ldots, A_i, \ldots, A_m) \tag{9}$$

In Equation (9), the test sequence is A, the common frame of the AS is m, and the feature vector of frame i is A_i. The sequence B is calculated as displayed in Equation (10).

$$B = (B_1, B_2, \ldots, B_j, \ldots, B_n) \tag{10}$$

In Equation (10), the template sequence is B, the common frame of the AS is n, and the feature vector of frame j is B_j. The ED $d(A_i, B_j)$ between the corresponding points of two ASs is Equation (11).

$$d(A_i, B_j) = \sqrt{\sum_{\omega=1}^{N} (A_{i\omega} - B_{j\omega})^2}, \quad 1 \leq \omega \leq N \tag{11}$$

In Equation (11), the feature value of the i-th frame of AS A is $A_{i\omega}$. The feature value of frame j in AS B is $B_{j\omega}$. The dimension of the AS is N. The regular path of DTW is

denoted as the line connecting the sequence alignment points, and the calculation of the regular path is Equation (12).

$$W = \{w_1, w_2, \ldots, w_k, \ldots, w_K\}, \quad \max(m, n) \leq k \leq m + n - 1 \tag{12}$$

In Equation (12), the planned path is W, and its k-th element is w_k. The dynamic regularization path requires to meet boundary constraints, starting from the bottom left corner and ending at the top right corner. Continuity constraint means that all points in an AS must appear in a regular path and be adjacent. Monotonicity constraint means that the regular path is monotonic on the timeline. The cumulative distance between feature vectors is Equation (13).

$$\gamma(i, j) = d(A_i, B_j) + \min\{\gamma(i-1, j-1), \gamma(i, j-1), \gamma(i-1, j)\} \tag{13}$$

In Equation (13), the cumulative distance $\gamma(i, j)$ between feature vectors is denoted as the ED sum between vectors and the nearest element distance. To improve the recognition rate, this study improved the cumulative distance formula, and the improved cumulative distance calculation is Equation (14).

$$\gamma'(i, j) = d(A_i, B_j) + \min\{\alpha\gamma(i-1, j-1), \beta\gamma(i, j-1), \theta\gamma(i-1, j)\} \tag{14}$$

In Equation (14), the improved cumulative distance is $\gamma'(i, j)$. The optimization coefficients are α, β, and θ, respectively. The intuition of adding optimization coefficients comes from the observation that DTW algorithm may encounter issues such as mismatches, high computational complexity, and low recognition rate when dealing with dynamic time warping problems. The improved DTW algorithm introduces optimization coefficients to constrain the search space, reduce computational complexity, improve computational power and recognition rate, and make the regular path closer to the diagonal, optimizing the matching effect of time series. To reduce computational complexity, a suitable threshold T is introduced to limit the cumulative distance. When the accumulated distance exceeds the threshold T, the calculation stops. When the accumulated distance is less than the threshold T, similarity matching continues. The target of action assessment analysis is to evaluate the quality of the AS to be evaluated by comparing it with standard actions. The traditional method relies on manual observation and experienced coaches, referees, etc. to identify differences in movements and score them, but is greatly influenced by subjective factors. To reduce subjective influence, this study adopted an action scoring method based on bone data. The DTW algorithm calculates the DTW distance between the test AS and the standard AS as a similarity assessment parameter. To reduce computational complexity, 8 joint angle features were selected as feature parameters. By selecting an action in the dataset as the evaluation object, using the action video data of students majoring in sports as the standard AS template, and selecting different samples as the test AS, multiple operations are performed to observe the distribution of joint angle distance. The evaluation formula for a certain sports action is Equation (15).

$$S_a = S_c - (d_1 - d_2) \times f_c \tag{15}$$

In Equation (15), the score for the angle feature is S_a, the score for angle allocation is S_c, the distance value and minimum value of DTW are d_1 and d_2, respectively, and the loss parameter is f_c.

$$S_{total} = \sum_{a=1}^{8} S_a \tag{16}$$

In Equation (16), the sum of the joint angle scores for the 8 actions is S_{total}. By using the DTW algorithm and defining a set of evaluation formulas, actions can be objectively evaluated and scored, helping to analyze and improve their execution. This method can

be applied to various scenarios for evaluating movements, such as sports training, and dance performances. The process of sports action recognition and evaluation based on improved DTW algorithm is shown in Figure 5. Firstly, by collecting and preprocessing sports action video data, joint angle features are extracted. Then, standardize the features to meet the requirements of the DTW algorithm. Next, a certain number of sports action videos are selected as the training set, and the KNN algorithm is used for training to obtain the prediction results of the classifier. Afterwards, use the DTW algorithm to calculate the distance between the test action sequence and the standard action sequence, and set a threshold T to calculate the score of the action based on the distance and threshold. Finally, based on the calculated action scores, the actions are objectively evaluated and scored, and this method is applied to sports training, dance performances, and other scenarios, optimized and adjusted according to actual needs.

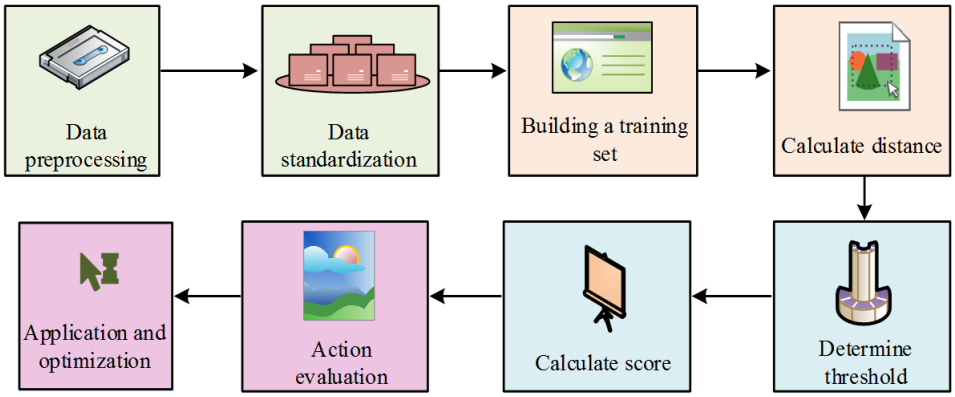
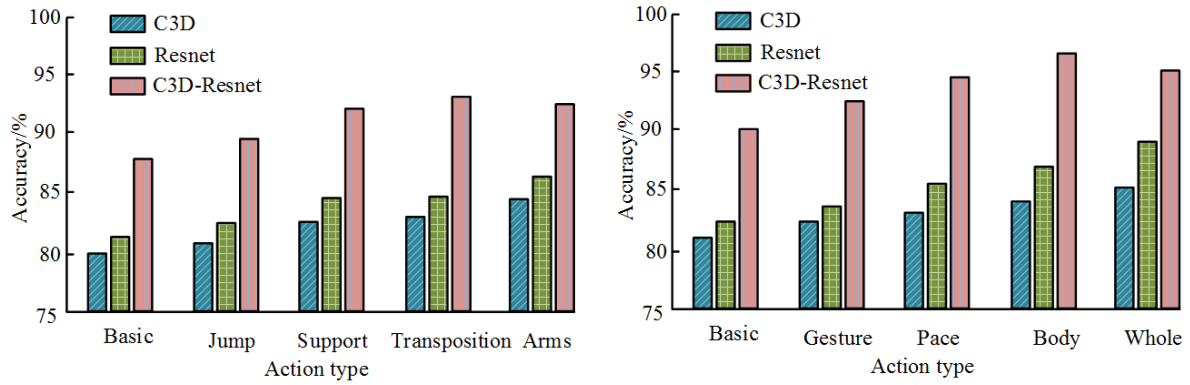


FIGURE 5. The process of sports action recognition and evaluation based on improved DTW algorithm

4. **Application Analysis of IEoSM Method Based on Improved DTW Algorithm.** This chapter first analyzes the application of the SAFE method, including the performance of the C3D-Resnet network on the Kinetics Dataset and CASIA TaiChi Dataset [22]. Secondly, the performance of the Attention-LSTM HMC extraction model was validated. Subsequently, the KNN-DTW algorithm was improved by dimensionality reduction and feature fusion, as well as selecting appropriate K values. Finally, the improved KNN-DTW algorithm was applied to sports action evaluation, and the actual application effect of this method was analyzed.

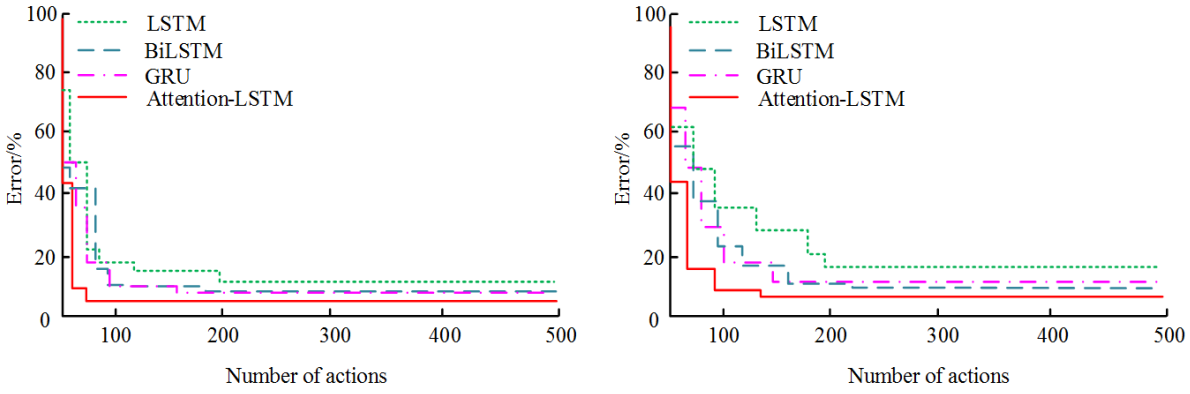
4.1. **Application analysis of SAFE method.** In the SAFE method experiment, a server equipped with a high-performance graphics card was used for hardware, with specific specifications of Intel Xeon processor, 32 GB memory, and NVIDIA GeForce RTX graphics card. The software uses the Ubuntu operating system and installs the CUDA acceleration library to support deep learning computation. The main programming language was Python and the deep learning framework utilized TensorFlow to build and train models. In addition, the OpenCV library is combined to handle the reading and preprocessing of video data. To test the C3D-Resnet's performance, C3D and Resnet were used as comparative methods in the experiment, and validated on the Kinetics Dataset and CASIA TaiChi Dataset sports action datasets. The Kinetics Dataset contains approximately 250000 videos, covering over 100 different sports categories with approximately 2000 samples per category. The CASIA TaiChi Dataset contains 22 Tai Chi movements, each with approximately 100 samples, for a total of approximately 2200 samples. The accuracy of different models in extracting environmental feature actions is Figure 6.



(a) Kinetics Dataset (b) CASIA TaiChi Dataset

FIGURE 6. The accuracy of different models in extracting environmental feature actions

Figure 6(a) shows the test results on the Kinetics Dataset sports dance dataset, with accuracy rates of 82.4%, 84.6%, and 90.5% for C3D, Resnet, and C3D-Resnet, respectively. Figure 6(b) shows the test results on the CASIA TaiChi Dataset sports Tai Chi dataset, with accuracy rates of 84.5%, 86.4%, and 94.7% for C3D, Resnet, and C3D-Resnet, respectively. In summary, the C3D-Resnet network has shown high accuracy on two different sports action datasets. In contrast, the accuracy of C3D and Resnet models is relatively low, especially on the CASIA TaiChi Dataset. Therefore, the C3D-Resnet network demonstrated better performance on sports dance and Tai Chi dataset. To verify the performance of the Attention-LSTM HMC extraction model, the experiment compared traditional LSTM with Gated Recurrent Unit (GRU) and the clustering based Bidirectional LSTM (BiLSTM) model proposed in 2023. The performance of different human body feature extraction models is Figure 7.



(a) Kinetics Dataset (b) CASIA TaiChi Dataset

FIGURE 7. The performance of different human feature extraction models

In Figure 7, through the comparison of four models, it is found that the Attention-LSTM HMC extraction model has the fastest convergence speed and the smallest error convergence value. In the Kinetics Dataset and CASIA TaiChi Dataset, Attention-LSTM converged at the 50th and 140th times, respectively, with error values of 7% and 9% after convergence. Compared to traditional LSTM, the convergence speed of Attention-LSTM has been improved by 51%, and the error value has been reduced by 8.5%. The results indicate that the improvement of this study is effective. The results indicate that

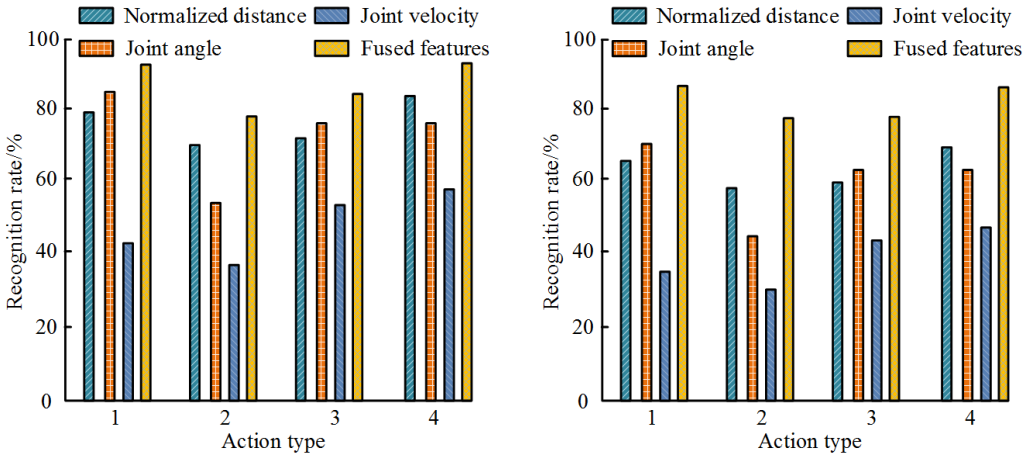
the model can learn and extract HMC faster, while also having higher accuracy and stability, providing strong support for further action recognition and behavior analysis. The experimental results of different combinations of EAC and HMC extraction models for SAFE are listed in Table 1.

TABLE 1. Experimental results of combining different environmental action features and human action feature extraction models

Dataset	EAC	HMC	Accuracy
Kinetics Dataset	C3D	Attention-LSTM	81.6%
	Resnet	Attention-LSTM	80.3%
	C3D-Resnet	BiLSTM	85.6%
	C3D-Resnet	Attention-LSTM	92.3%
CASIA TaiChi Dataset	C3D	Attention-LSTM	84.1%
	Resnet	Attention-LSTM	85.2%
	C3D-Resnet	BiLSTM	86.6%
	C3D-Resnet	Attention-LSTM	94.7%

In Table 1, among the different combinations of models, the accuracy of the C3D-Resnet and Attention-LSTM combination models on the Kinetics Dataset and CASIA TaiChi Dataset is 92.3% and 94.7%, respectively. The results showed that the combination model of C3D Resnet and Attention-LSTM achieved high accuracy in action recognition tasks, demonstrating the importance of considering EAC and HMC comprehensively. This provides strong reference and guidance for further research on action recognition and behavior analysis, which can help better understand and utilize information in sports actions.

4.2. **Application analysis of IEoSM method.** The experiment analyzed the KNN-DTW's performance by studying the collected Tai Chi data information. The results of different feature recognition are exhibited in Figure 8.

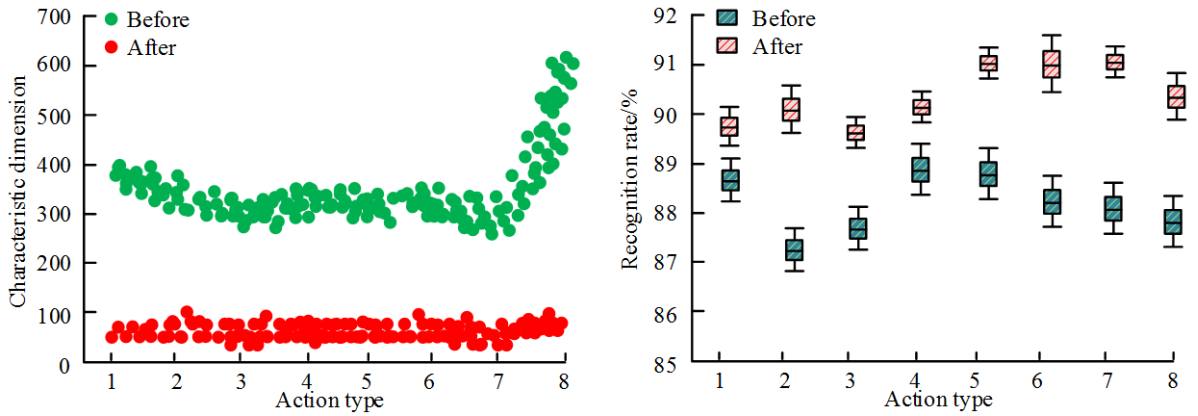


(a) Low difficulty movements (b) High difficulty movements

FIGURE 8. Different feature recognition results

The results in Figure 8 indicate that the recognition rate of fused features outperforms that of single features, proving the effectiveness of fused features. Meanwhile, the classification performance of joint velocity features is poor, which may be due to the small difference in the position changes of joint points between frames of different actions and the

unequal time interval between two frames of different actions. To reduce data redundancy, Principal Component Analysis (PCA) can be used to reduce the dimensionality of fused features and extract the main components. The result of Feature Data Dimensionality Reduction (FDDR) is Figure 9.



(a) Characteristic dimension (b) Recognition rate

FIGURE 9. Result of feature data dimensionality reduction

In Figure 9, before and after dimensionality reduction, the feature data dimension decreased from $[289, 678]$ to within the range of $[17, 109]$, and the average action recognition rate increased from 88.3% to 91.2%. This indicates that FDDR can significantly reduce the dimensionality of feature vectors while maintaining a high recognition rate. By reducing the data dimension, redundant information can be effectively reduced, and the expression ability and classification effect of features can be improved. Therefore, FDDR is an effective method that can further optimize the performance and effectiveness of action recognition. To verify the KNN-DTW's effectiveness with improved dimensionality reduction and fusion features, this study used data from three datasets as experimental samples. Table 2 provides information on the action recognition dataset.

TABLE 2. Action recognition dataset information

Dataset	Action number									
	1	2	3	4	5	6	7	8	9	10
Self-made	300	300	300	300	300	300	300	300	/	/
Kinetics Dataset	15	40	50	30	40	30	35	25	32	25
CASIA TaiChi Dataset	500	530	480	520	550	470	/	/	/	/

In Table 2, the action types and quantities of the three experimental datasets are different, which can be used to comprehensively evaluate algorithm performance. The relationship between the parameters determined by K-fold crossover and algorithm performance is Figure 10.

In Figure 10, as the K value changes, the recognition rate of the improved KNN-DTW also changes. This indicates that the K value selection is important in the performance of the algorithm. A smaller K value may lead to over-fitting, while a larger one may cause under-fitting. When conducting K-fold cross validation, the recognition rate of the algorithm reached its highest point when the K value was set to 5. Therefore, the K value of the algorithm will be set to 5 in subsequent experiments. The comparison of action evaluation results of the improved KNN-DTW algorithm on three datasets is shown in Figure 11.

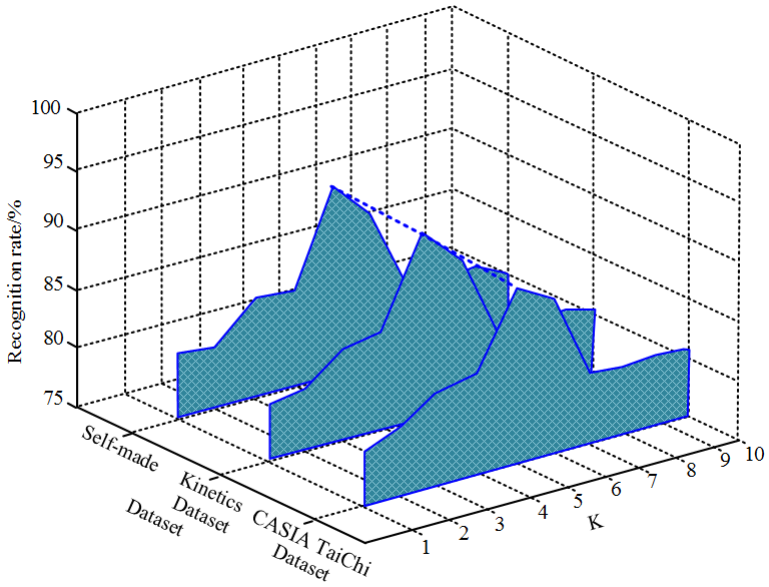
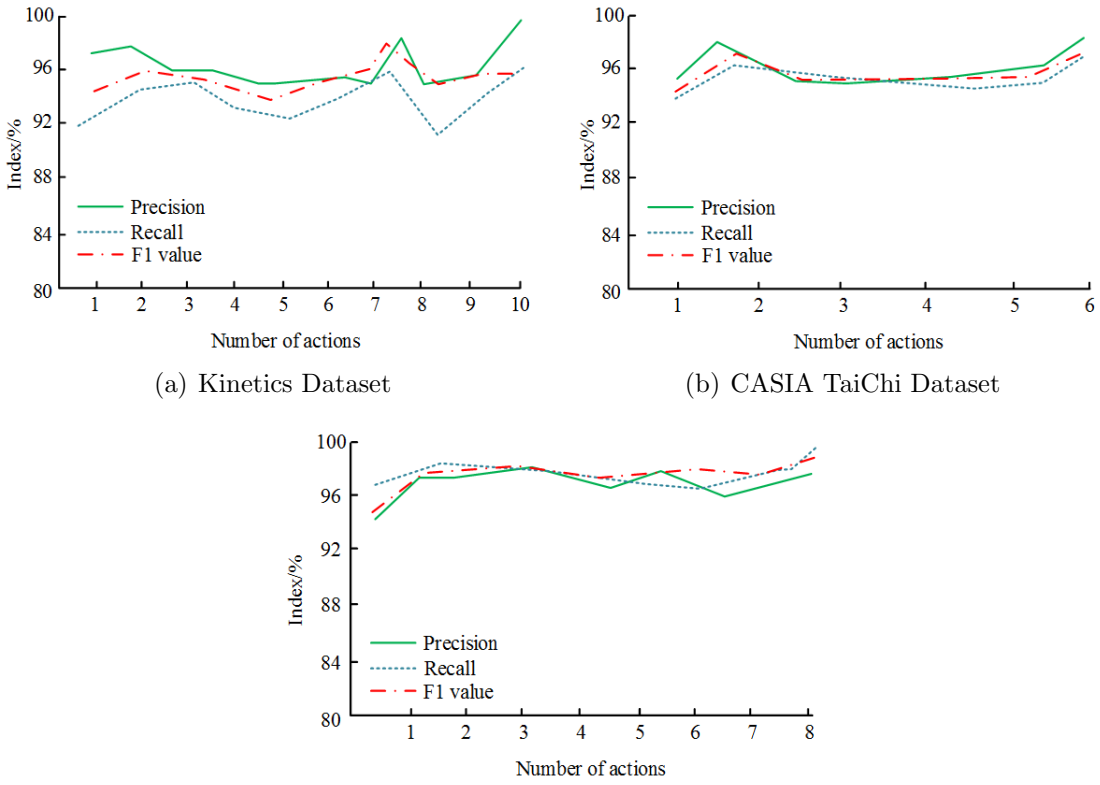


FIGURE 10. The relationship between parameters determined by K-fold crossover and algorithm performance



(a) Kinetics Dataset

(b) CASIA TaiChi Dataset

(c) Self-made

FIGURE 11. Improved KNN-DTW algorithm for action evaluation results on three datasets

In Figure 11, the improved KNN-DTW algorithm has an F1 value of approximately 97.1% on three datasets. Therefore, the improved KNN-DTW algorithm shows high accuracy and reliability in evaluating actions on these three datasets. After substituting the action evaluation formula, the difference between the sports action evaluation results

and the scores of professional coaches is less than 1 point, indicating that the practical application effect of this method is relatively ideal. To verify the effectiveness of the improved KNN-DTW algorithm, traditional DTW algorithm and Convolutional Neural Network (CNN) were compared in the experiment. Simultaneously, the state-of-the-art Faster R-CNN and YOLOv5 algorithms are used as a comparison. The comparison of the effects of different algorithms is shown in Table 3. In Table 3, it can be seen that the improved KNN-DTW algorithm is significantly better than DTW and CNN algorithms in different datasets. This indicates that the improved KNN-DTW algorithm has high practical value in motion action classification tasks. Meanwhile, the F1 value of the improved KNN-DTW algorithm has been improved on the basis of Faster R-CNN and YOLOv5, indicating that this improvement has higher accuracy and performance in processing object detection tasks. The improved KNN-DTW algorithm can better explore potential patterns and features in time series data, thereby improving the performance of object detection tasks.

Table 3. Comparison of the effects of different algorithms

Dataset	Algorithm	F1 value
Kinetics Dataset	KNN-DTW	96.7%
	DTW	93.2%
	CNN	92.1%
	Faster R-CNN	95.2%
	YOLOv5	96.1%
CASIA TaiChi Dataset	KNN-DTW	96.3%
	DTW	93.7%
	CNN	93.6%
	Faster R-CNN	95.7%
	YOLOv5	95.9%
Self-made	KNN-DTW	98.2%
	DTW	92.7%
	CNN	93.5%
	Faster R-CNN	97.1%
	YOLOv5	97.5%

5. **Conclusion.** Motor performance evaluation is an important means of providing feedback and evaluation to athletes during training and competition. It can help athletes understand their physical condition, adjust training plans, and improve training effectiveness. To accurately identify and evaluate the performance of athletes in competitions, this study applied the DTW algorithm to IEoSM and combines CNN for feature extraction. On the sports dance dataset, the accuracy of C3D, Resnet, and C3D-Resnet were 82.4%, 84.6%, and 90.5%, respectively. On the CASIA TaiChi Dataset, the accuracy of C3D, Resnet, and C3D-Resnet were 84.5%, 86.4%, and 94.7%, respectively. Compared to traditional LSTM, the convergence speed of Attention-LSTM has been improved by 51%, and the error value has been reduced by 8.5%. The F1 value of the improved KNN-DTW algorithm on three datasets was about 97.1%. After substituting the action evaluation formula, the difference between the sports action evaluation results and the scores of professional coaches was less than 1 point. The outcomes proved that the designed IEoSM grounded on the improved DTW is an effective method for identifying and evaluating the movement performance of athletes. In practical applications, the DTW algorithm also needs to consider its operability, such as parameter settings and computational efficiency.

Future research can explore how to apply the DTW algorithm to more complex sports action recognition and evaluation to improve its accuracy and robustness.

REFERENCES

[1] M. Riahi, M. Eslami, S. H. Safavi and F. T. Azar, Human activity recognition using improved dynamic image, *IET Image Processing*, vol.14, no.5, pp.3223-3231, 2020.

[2] J. Zhang, Y. Cao and Q. Wu, Vector of locally and adaptively aggregated descriptors for image feature representation, *Pattern Recognition*, vol.116, no.4, pp.107952-107962, 2021.

[3] P. Sharma and R. S. Anand, Depth data and fusion of feature descriptors for static gesture recognition, *IET Image Processing*, vol.14, no.5, pp.909-920, 2020.

[4] O. Kechagiasstamatis, N. Aouf and M. Richardson, Performance evaluation of single and cross-dimensional feature detection and description, *IET Image Processing*, vol.14, no.10, pp.2035-2051, 2020.

[5] J. Cai, J. Hu, S. Li, J. Lin and J. Wang, Combination of temporal-channels correlation information and bilinear feature for action recognition, *IET Computer Vision*, vol.14, no.8, pp.634-641, 2020.

[6] H. Zhao, W. Xue, X. Li, Z. Gu, L. Niu and L. Zhang, Multi-mode neural network for human action recognition, *IET Computer Vision*, vol.14, no.8, pp.587-596, 2020.

[7] Z. Ren, Q. Zhang, P. Qiao, M. Niu, X. Gao and J. Cheng, Joint learning of convolution neural networks for RGB-D-based human action recognition, *Electronics Letters*, vol.56, no.21, pp.1112-1115, 2020.

[8] B. Chen, H. Tang, Z. Zhang, G. Tong and B. Li, Video-based action recognition using spurious-3D residual attention networks, *IET Image Processing*, vol.16, no.11, pp.3097-3111, 2022.

[9] F. Xue, H. Ji and W. Zhang, Mutual information guided 3D ResNet for self-supervised video representation learning, *IET Image Processing*, vol.14, no.13, pp.3066-3075, 2020.

[10] P. Gao, D. Zhao and X. Chen, Multi-dimensional data modelling of video image action recognition and motion capture in deep learning framework, *IET Image Processing*, vol.14, no.7, pp.1257-1264, 2020.

[11] X. Yu, Evaluation of training efficiency of table tennis players based on computer video processing technology, *Optik*, vol.273, no.1, pp.170404-170408, 2023.

[12] T. T. Nguyen, D. T. Pham, H. Vu and T. Le, A robust and efficient method for skeleton-based human action recognition and its application for cross-dataset evaluation, *IET Computer Vision*, vol.16, no.8, pp.709-726, 2022.

[13] Y. Zhao, M. Guo, X. Sun, X. Chen and F. Zhao, Attention-based sensor fusion for emotion recognition from human motion by combining convolutional neural network and weighted kernel support vector machine and using inertial measurement unit signals, *IET Signal Processing*, vol.17, no.4, pp.12201-12212, 2023.

[14] J. Xu, H. H. Li and S. Hou, Attention-based gait recognition network with novel partial representation PGOFI based on prior motion information, *Digit. Signal Process.*, vol.133, no.1, pp.103845-103849, 2023.

[15] L. Wu, Z. Li, Y. Xiang, M. Jian and J. Shen, Latent label mining for group activity recognition in basketball videos, *IET Image Processing*, vol.15, no.14, pp.3487-3497, 2021.

[16] Y. Li, K. Li and X. Wang, Recognizing actions in images by fusing multiple body structure cues, *Pattern Recognition*, vol.104, no.10, pp.107341-107352, 2020.

[17] S. Ali and N. Bouguila, Multimodal action recognition using variational-based Beta-Liouville hidden Markov models, *IET Image Processing*, vol.14, no.17, pp.4785-4794, 2020.

[18] X. Li, M. Xie, Y. Zhang and G. Ding, Dual attention convolutional network for action recognition, *IET Image Processing*, vol.14, no.6, pp.1059-1067, 2020.

[19] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu and C. Xu, GhostNet: More features from cheap operations, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol.1, no.1, pp.1580-1589, 2020.

[20] R. Kumar and S. Kumar, Survey on artificial intelligence-based human action recognition in video sequences, *Optical Engineering*, vol.62, no.2, pp.23102-23122, 2023.

[21] J. Purohit and R. Dave, Leveraging deep learning techniques to obtain efficacious segmentation results, *Archives of Advanced Engineering Science*, vol.1, no.1, pp.11-26, 2023.

[22] C. Hebbi and H. Mamatha, Comprehensive dataset building and recognition of isolated handwritten Kannada characters using machine learning models, *Artificial Intelligence and Applications*, vol.1, no.3, pp.179-190, 2023.

Author Biography

Yuli Hu is a master's degree holder. Now he is an associate professor, and director of the Military Sports Teaching and Research Office of the Aviation Fundamentals College of the Naval Aviation University. He has published multiple articles, like *Research and application analysis of sports group training theory* in *Ludong University Journal* (Natural Science Edition) in 2007, and *Research on the lack and countermeasures of water rescue education* in Colleges and Universities Education and Teaching Forum 2017.

Di Liu works as a professor, and director of the Teaching and Research Office for Automatic Detection Technology at the School of Engineering, Naval Aviation University. He has published multiple articles in electronic design engineering, automation, and instrumentation, and has been awarded by the National Natural Science Foundation of China multiple times.