Journal of Ambient Intelligence and Humanized Computing (2024) 15:3317–3330
https://doi.org/10.1007/s12652-024-04809-8

ORIGINAL RESEARCH

Research on badminton take-off recognition method based on improved deep learning

Lu Lianju[1] · **Zhang Haiying**[2]

Received: 23 October 2023 / Accepted: 28 April 2024 / Published online: 28 May 2024

Abstract
Because of the fast take-off speed of badminton, a single action recognition method can't quickly and accurately identify the action. Therefore, a new badminton take-off recognition method based on improved deep learning is proposed to capture badminton take-off accurately. Collect badminton sports videos and get images of athletes' activity areas by tracking the moving targets in badminton competition videos. The static characteristics of badminton players' take-off actions are extracted from the athletes' activity areas' images using 3D ConvNets. According to the human joint points in the badminton player's target tracking image, the human skeleton sequence is constructed by using a 2D coordinate pseudo-image and 2D skeleton data design algorithm, and the dynamic characteristics of badminton take-off action are extracted from the human skeleton sequence by using LSTM (Long-term and Short-term Memory Network). After the static and dynamic features are fused by weighted summation, badminton take-off feature fusion results are input into a convolutional neural network (CNN) to complete badminton take-off recognition. The CNN pool layer is improved by adaptive pooling, and the network convergence is accelerated by combining batch normalization to further optimize the recognition results of badminton take-off. Experiments show that the human skeleton model can accurately match human movements and assist in extracting action features. The improved CNN has greatly improved the accuracy of recognition of take-off actions. When recognizing real images, it can accurately identify human movements and judge whether there is a take-off action.

Keywords Human skeleton sequence · Convolutional neural network · Static characteristics · Dynamic characteristics · Pool layer · Feature fusion

1 Introduction

Motion recognition is the core problem of video analysis and understanding, and it is also a hot and difficult problem in computer vision (Arashpour et al. 2022) and pattern recognition. Its research spans computer vision, pattern recognition, artificial intelligence (Shin 2023), image processing, machine learning and other disciplines (PlÖtz 2021). The goal of motion recognition is to classify the specific actions of people in video (Leibovich et al. 2020) and then provide a basis for video content analysis and semantic understanding. Specifically, relevant personnel can extract the information related to behaviors and actions in the video data by analyzing it, thus establishing a bridge between the underlying data and the semantic understanding of high-level behaviours. Deep learning technology is gradually emerging (Kesavavarthini et al. 2023), and it has broken through the performance limit of traditional methods in various fields of image processing (Khaddam et al. 2022). Deep learning can automatically learn the features required by a specific application from a large number of original data (Tang et al. 2022), and it does not need much professional and complicated feature design. Compared with other motion recognition technologies, deep learning takes the pixels of the original image as input (Davtalab et al. 2022; Jurado et al. 2022) and simulates the processing of natural images by the nervous system of the human brain with a multi-layer network structure, thus achieving the purpose of understanding the semantic content of images. In competitive sports, professional athletes and their teams need to replay the video

✉ Zhang Haiying
270971194@qq.com

[1] College of General Education, Liaoning University of ForeignTrade and Economics, Dalian 116052, China

[2] Sports Education, Liaoning University of International Business and Economics, Dalian 116052, China

of each game, and the adoption of motion recognition can effectively reduce the workload of the coaching team to analyze the replay and learn more about the badminton game process, discover technical details and extract wonderful clips. Therefore, the related human motion recognition methods have attracted extensive attention from scholars.

Pau Climent-Pérez et al. developed an improved action recognition method with separable spatiotemporal attention based on bone and video preprocessing (Climent-Perez et al. 2021), which used a separable spatiotemporal attention network combined with view-invariant normalization of full activity clipping and bone posture data of RGB data to form a recognition network, which can effectively recognize various actions of the human body. Avola et al. studied the action and interaction recognition method based on multi-view representation of manual low-level skeleton features (Avola et al. 2022). Using multi-view representation learning and selecting whole-body manual features only through psychology and close-range research, two-dimensional skeleton data were extracted from RGB video sequences to obtain various low-level skeleton features, that is, multi-views. The codebook related to conditions was generated by the visual bag clustering method. Each emotional action and interaction in the video could be expressed as the frequency histogram of codewords. The training samples were used to calculate and store them in the database. Mokari M et al. studied the method of identifying actions from 3D bone data using body states (Mokari et al. 2017), defined body states, and modelled each action as a sequence of these states. In the learning stage, Fisher linear discriminant analysis (LDA) is used to construct a discriminant feature space to discriminate the body state. In addition, Mahalanobis distance is used as an appropriate distance measure to classify behavior states. Then, the Hidden Markov Model (HMM) simulates the time transition between body states in each movement. The video motion recognition based on a pseudo-3D residual attention network (Chen et al. 2022), researched by Chen B et al., forms a pseudo-3D residual attention network (S3D RANs) by improving the 3D convolutional neural network (3D CNN). The improved network takes advantage of the advantages of 2D convolutional neural network (2D CNN) and 3D CNN and applies 2D CNN to the frame of the single view of volume video data. Directly learn the characteristics of time motion, and the residual unit adopts the sub-module of view and channel attention mechanism to learn the importance of each view to action recognition and guide the network to pay attention to more useful information for action recognition to complete the action recognition of people. The multi-modal human behavior recognition of autonomous system based on ambient intelligence (Jain et al. 2023) researched by Envelope JA et al. uses Bi-CRNN for feature extraction and random forest classification and uses automatic fusion technology to improve the fusion of data from various sensors, making the result of action recognition more accurate. Although the above method can complete motion recognition, it also has some disadvantages, such as when using a single convolutional neural network (Ying et al. 2020), it is difficult to capture high-intensity and fast sports, the accuracy and speed of feature extraction are poor, the accuracy of motion recognition is poor, and the calculation consumes more resources.

Deep learning (Sun et al. 2021) methods include CNN, LSTM, etc. The convolutional neural network has great advantages in image recognition and feature extraction, such as high efficiency, adaptive learning and wide application. Therefore, this paper proposes a badminton take-off recognition method based on improved deep learning. Three deep learning methods, 3D ConvNets network, LSTM and CNN, are comprehensively used to accurately identify the take-off action in badminton by extracting features and recognizing them.

1.1 Badminton take-off action recognition

The deep learning method includes many categories, considering that the recognition of badminton take-off action needs to adapt to different modes of video information and fully tap the static and dynamic characteristics of badminton take-off action. Therefore, when designing the badminton take-off recognition method, this paper combines the advantages of using the 3D ConvNets network, LSTM and CNN to design an improved deep learning method to realize the accurate recognition of the take-off action. The overall frame is shown in Fig. 1.

For the static characteristics of badminton take-off, based on the badminton competition video, the image of athletes' activity area is obtained by tracking the moving targets in the badminton competition video, and the static characteristics are obtained by combining 3D ConvNets. For the dynamic characteristics of badminton take-off, based on the images of athletes' activity areas, the data set containing the human skeleton sequence is obtained by constructing the human skeleton sequence of badminton players, and the dynamic characteristics are extracted from it by LSTM model. After the weighted summation method fuses the static and dynamic characteristics of badminton take-off (Wen et al. 2020), the fused characteristics are input into improved CNN to realize badminton take-off recognition.

1.2 Badminton player target tracking image generation

In the video of badminton, the shooting position is usually non-static. Therefore, there are many moving pictures

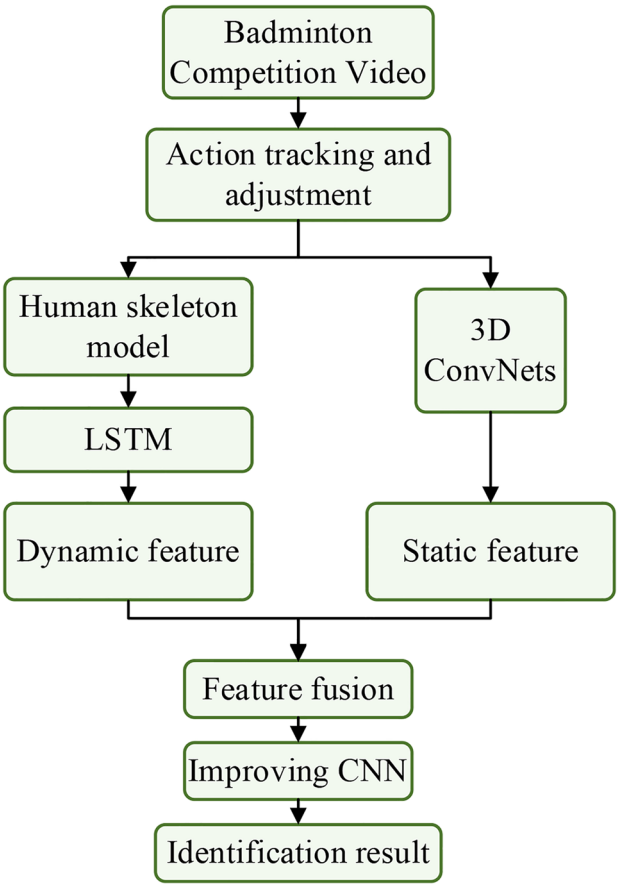


Fig. 1 Multi-modal identification framework

in badminton sports videos. The movement information of badminton players cannot be accurately reflected due to the camera's movement. Therefore, a badminton action recognition method is proposed, which obtains the badminton player's activity area by tracking the moving target, and adjusts the player's image in this area to offset the camera motion in the video. The target area is tracked in the horizontal and vertical directions of regional expansion symmetry to adjust the image containing badminton moving targets. Calculate the coordinates of the "center of mass" of the target area using formula (1) (m_x, m_y), and move the center of the tracking area to the "center of mass" coordinate to complete the badminton player's target tracking image generation:

$$\begin{cases} m_x = \frac{\sum_{x\in R}\sum_{y\in R} x \times f(x,y)}{\sum_{x\in R}\sum_{y\in R} f(x,y)} \\ m_y = \frac{\sum_{x\in R}\sum_{y\in R} y \times f(x,y)}{\sum_{x\in R}\sum_{y\in R} f(x,y)} \end{cases} \tag{1}$$

In the formula, R is a tracking area for badminton players. $f(x, y)$ expresses the gray value of pixels in R.

After the above processing, the target tracking adjustment image sequence can eliminate the influence of camera movement and always follow the movement of athletes.

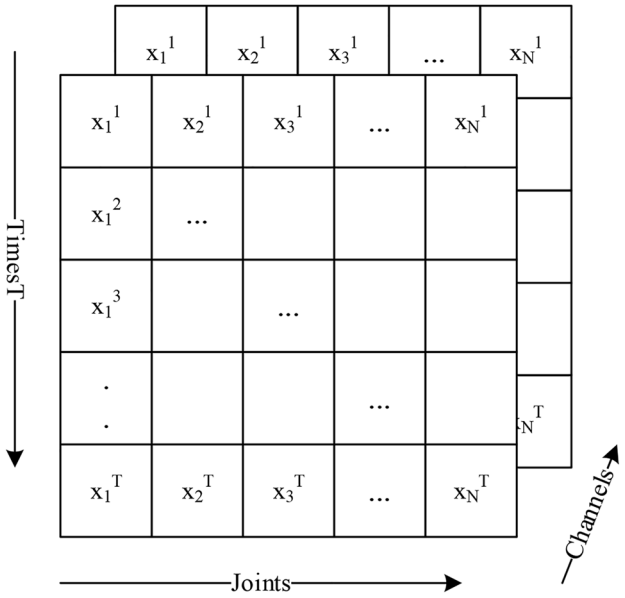


Fig. 2 Pseudo-image of node coordinates

Therefore, the adjusted image sequence only contains the movement caused by the player's limbs and racket and does not reflect the camera's movement in the original video.

1.3 Human skeleton sequence of badminton players

The human skeleton has been widely used in behavior recognition tasks (Peng et al. 2021), in which the coordinates of human joints are usually constructed as joint sequences, pseudo-images or graphs. The human skeleton sequence is of great significance to behavior recognition. This paper introduces a new bone shape, bone edge movement, by giving the coordinates of joint points, which helps learn effective features by exploring the movement of body parts.

In badminton, take-off recognition (Thangarajan et al. 2021), skeleton data, that is, the coordinates of badminton players' joints, are usually constructed in pseudo-images, as shown in Fig. 2, which are 2D coordinate pseudo-images.

However, the body of badminton players contains prior knowledge of human topological structure, which is difficult to learn for deep neural networks. For the convenience of description, the joint points are coded in sequence. To simplify, this paper designs an algorithm based on 2D bone data, which can be easily extended to 3D bones. In this paper, firstly, the OpenPose algorithm is used to extract the set of human joints in badminton players' target tracking images $\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$, and then, according to the joint point (x_i, y_i) and (x_j, y_j) define bone edges e_{ij}, constitute the human body topology. Therefore, defined in time t, the motion mode of the bone edge on the, that is, the rotation angle and moving distance of e_{ij}, are

$$cos\Delta\theta_i^t = \frac{a_i^t \cdot a_i^{t+1}}{m_x \cdot m_y} \tag{2}$$

$$a_i^t = \left[x_j^t - x_i^t, y_j^t - y_i^t\right] \tag{3}$$

In the formula, $cos\Delta\theta_i^t$ is the rotation angle of e_{ij}.

Since limbs can be roughly regarded as rigid bodies, this paper can pass through joint points (x_i, y_i) and (x_j, y_j) determine the moving distance.

$$\Delta l_i^t = cos\Delta\theta_i^t \sqrt{\left(x_i^{t+1} - x_i^t\right)^2 + \left(y_i^{t+1} - y_i^t\right)^2} \tag{4}$$

In the formula, Δl_i^t is the moving distance of e_{ij}.

Combined with the static coordinates of joint points and $\Delta\theta$ and Δl to define the motion modes of bone edges as different dimensions of pseudo-images. At this time, a characteristic matrix $I \in R^{T\times N\times C}$ is obtained, in which T represents the number of video frames, N indicate the number of joint points, C represents the dimension of the joint point coordinates, such as 2D bone data $C = 4$, and 3D bone data $C = 5$. In this paper, nine bone edges are used, namely $e_{2,1}$, $e_{3,4}$, $e_{4,5}$, $e_{6,7}$, $e_{7,8}$, $e_{9,10}$, $e_{12,13}$ and $e_{13,14}$ corresponding to the neck, arms and legs of the badminton player's body, that is, the bone structure obtained as shown in Fig. 3.

1.4 Badminton take-off feature fusion based on deep learning

This paper's badminton take-off feature extraction is completed based on two modes: RGB video stream and human skeleton points. RGB information stream contains the global information of environment and character movement, and 3DConvNets can be used to mine the spatial and temporal information of video. The skeleton sequence contains the movement information of the key points of badminton players in the video, so it is more accurate to mine the movement information of badminton players.

1.4.1 Badminton take-off action static feature extraction

For the traditional RGB video stream of badminton take-off, 3D ConvNets can mine not only the spatial characteristics of a single video frame but also the temporal characteristics of a short sequence. 3D ConvNets network superposes multiple video frames of badminton take-off action to form a cube. Through 3D convolution operation, each feature map will be partially connected with multiple consecutive frames in the upper layer, thus retaining the time sequence information (Totaro et al. 2020).

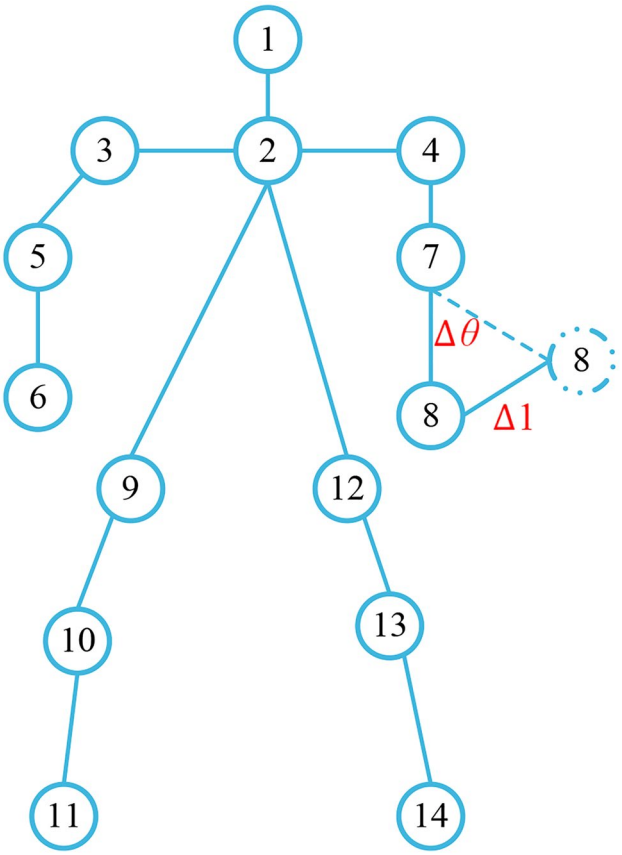


Fig. 3 Code of the node

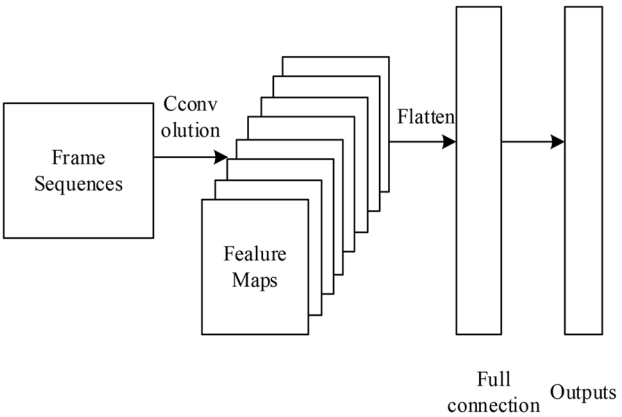


Fig. 4 Processing of continuous video frames using 3D ConvNets

The 3D ConvNets network pre-trained on the Sports-1M database is used, and the network structure layer is shown in Fig. 4.

Change the last fully connected output of the original 3D ConvNets network to the dimension corresponding to the number of data sets. The processing of continuous video frames using 3D ConvNets is shown in Fig. 4.

1.4.2 Badminton take-off action dynamic feature extraction

The LSTM recurrent neural network is used to process the human skeleton sequence data set in the skeleton model, and the dynamic characteristics of the take-off action of hairball are obtained.

The LSTM cycle unit structure uses three gating mechanisms (Kumar et al. 2022), as shown in Fig. 5, to effectively transmit and express the information in the long time series of bone point information without causing the useful information long ago to be forgotten, thus realizing the preservation and transmission of context information in a large range. In this paper, the Sigmoid function is used as the activation function of the gate control unit (Rico 2022), and the Tanh function is used as the activation function of the unit state and output.

The detailed working principle of the LSTM unit is as follows (Harvat et al. 2022):

The forgetting gate indicates how much information about the last time point should be forgotten, which determines how much unit status information is at the previous moment C_{t-1} and how much of the skeleton sequence is saved to the unit state C_t at the current moment. Connect from the previous moment of hidden state h_{t-1} and the current moment of input x_t, to form a new feature vector, and multiplied with the weight parameter W_f after input to sigmoid activation function, the output vector f_t with the corresponding elements with C_{t-1} multiplication operation: $C_{t-1} \times f_t$ to determine the previous moment badminton player skeleton sequence shallow feature information unit state C_{t-1} how many badminton player skeleton sequence shallow feature information is added to the current unit state C_t, the closer the element in f_t is to 0, indicates that the more the shallow feature information of the badminton player skeleton sequence in C_{t-1} is forgotten, while on the contrary, the closer the element in f_t is to 1, the more the shallow feature information of the badminton player skeleton sequence in C_{t-1} is retained. f_t is calculated as shown in Formula (5):

$$f_t = \sigma\left(\frac{W_f}{\Delta l_i^t}[h_{t-1}, x_t] + b_f\right) \tag{5}$$

In the formula, W_f and b_f represent the weight parameter and bias term of sigmoid activation function in forgetting gate, respectively.

The input gate indicates how much shallow feature information of the input badminton player skeleton sequence should be remembered at the current time point, which determines the shallow feature information of the input badminton player skeleton sequence at the current time unit x_t and how many are saved to the current cell state C_t . The shallow feature information of the candidate badminton player skeleton sequence $\overline{C}_t$ is determined by the tanh activation function, multiplying the corresponding elements of the decision vector I_t and candidate information $\overline{C}_t : \overline{C} \times I_t$ to determine how much $\overline{C}_t$ is added to the cell state C_t The calculation of I_t and $\overline{C}_t$ are:

$$I_t = \sigma\left(W_I \cdot [h_{t-1}, x_t] + b_I\right) \tag{6}$$

$$\overline{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_c\right) \tag{7}$$

In the formula, W_I and W_C respectively represent the weight parameters corresponding to sigmoid and tanh in the input gate (Sun et al. 2020), b_I and b_c represent the corresponding offset. Cell state at the current moment C_t is:

$$C_t = C_{t-1} \times f_t + \overline{C}_t \times I_t \tag{8}$$

The output gate indicates how much shallow feature information of the badminton player skeleton sequence should be output from the unit state information at the current time point, which determines the unit state at the current time point C_t and how much output is to the hidden state of the cell h_t. The decision vector of o_t and the hidden state of the cell h_t of the cell state C_t is:

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{9}$$

$$\overline{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_c\right) \tag{10}$$

Among them, W_o and b_o represent the weight parameter and bias term of the sigmoid activation function in the output gate, respectively.

The process of human skeleton feature extraction using LSTM is shown in Fig. 6.

The LSTM network is used to extract features from the badminton player's human skeleton sequence, and

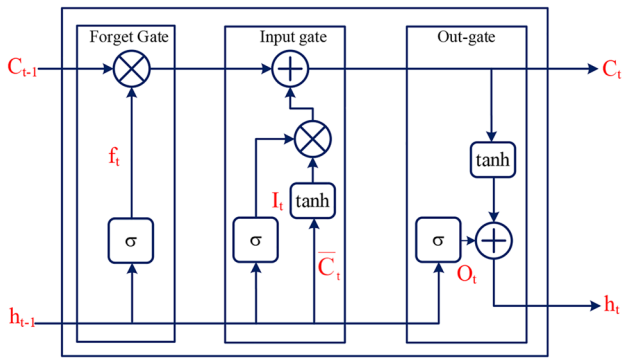


Fig. 5 LSTM cell structure

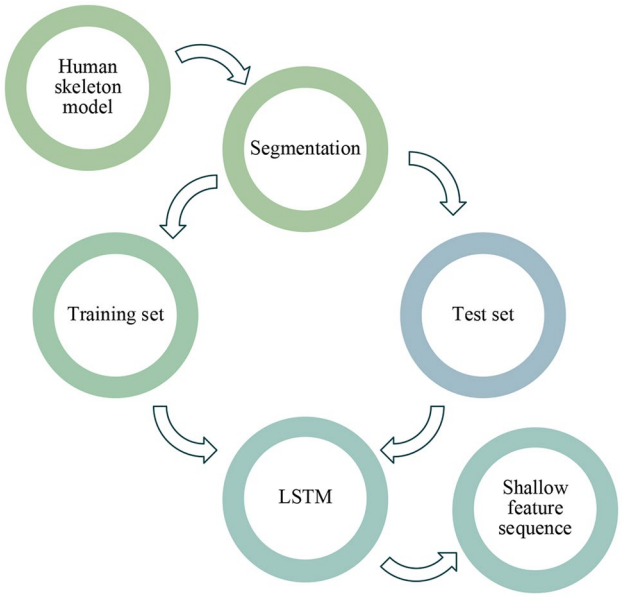


Fig. 6 Human skeleton feature extraction process

the three-dimensional position information of several joint points corresponding to the posture at each moment is obtained from the badminton player's human skeleton sequence, which is used to construct the shallow features of the skeleton sequence, and the shallow features at T moment are expressed as f^t, the shallow feature set containing n frames of badminton players' human skeleton sequence is: $V = \{f^t | t = 1, 2, \cdots, N\}$.

Because the shallow features of badminton players' human skeleton sequence still belong to time series (Ozdemir et al. 2021), the LSTM network can be used to learn the corresponding time dynamic information. In the time dynamic feature extraction layer composed of several LSTM units with the same structure, each LSTM unit is only responsible for processing a feature sequence, which corresponds to a coordinate component of a badminton player's skeleton joint in this method, so that each LSTM unit can learn the corresponding time dynamic information. The shallow feature sequence is constructed by the human skeleton sequence of badminton players, and the *m* shallow feature sequence in *t* time the value is used $s(m, t)$ to show. Then, $s(m, :)$ is the *m* shallow features of the human skeleton sequence of badminton players are input by the LSTM unit.

The fully connected neural network and the Softmax layers realize the classifier used in this paper. The input of the fully connected neural network layer is the linear combination of human skeleton features output by all the LSTM units in the LSTM layer at the last time step. This layer combines the different time dynamic information learned in each LSTM unit and outputs the final badminton take-off dynamic feature extraction result.

1.4.3 Take-off action feature fusion

After obtaining the static characteristics and dynamic characteristics of badminton players' take-off, better results can be achieved by integrating these two characteristics. The specific methods are as follows:

Hypothetical weight vector $Z = (a_1, a_2, \cdots, a_m)$, the characteristics of badminton players' take-off movements are fused, and the formula is:

$$Y_{fusion} = ZY^T = h_t\left(a_1 Y_1^t + a_2 Y_2^t + \cdots + a_m Y_m^t\right) \tag{11}$$

In the formula, Y_{fusion} represents the characteristics of fusion *m* represents the number of features extracted by the two algorithms, Y_m^t represents dynamic and static characteristics.

1.5 Badminton take-off recognition based on improved CNN

1.5.1 CNN network architecture design

Take-off characteristics of badminton after integration Y_{fusion} input it into CNN network to identify the take-off action of badminton. CNN network structure is shown in Fig. 7.

As can be seen in Fig. 7, the CNN network structure is mainly divided into five layers: input layer, convolution layer, sub-sampling layer (pooling layer), full connection layer and output layer.

Input layer: The take-off characteristics of badminton after fusion.

Convolution layer: The convolution layer uses multiple convolution kernels to extract depth features from badminton take-off feature fusion image. Let the original fused badminton take-off action feature image, be $X = (x_{i_1 j_1})_{\alpha\times\alpha} = \sum Y_{fusion}$, convolution kernel is $W = (w_{i_2 j_2})_{\beta\times\beta}$, the convolution kernel dimension is d, convolution kernel moving step is *e*. Perform convolution operation to X and W, and the output is $Y' = (y_{i_3 j_3})_{\gamma\times\gamma}$, then $y_{i_3 j_3}$ is the sum of the products of all the elements of and the elements in the same position of the convolution kernel of X with the abscissa from $(1 + e(i_3 - 1))$ to $(d + e(i_3 - 1))$ and the ordinate from $(1 + e(j_3 - 1))$ to $(d + e(j_3 - 1))$, and then the activation function Relu. Relu can zero the negative value of the input matrix, and return the positive value according to the original value, that is

$$Relu(x) = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases} \tag{12}$$

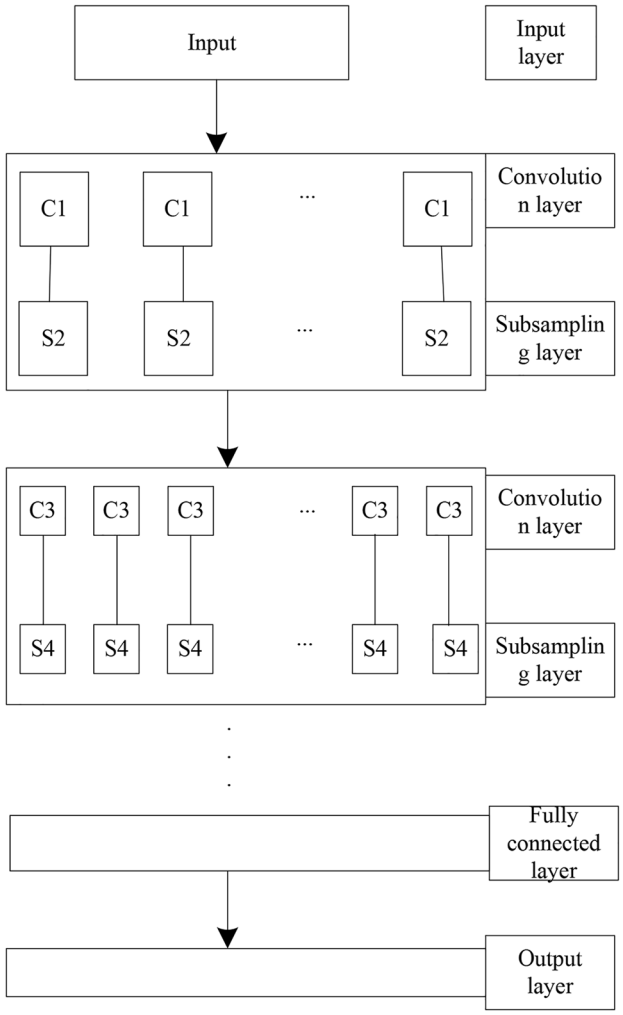


Fig. 7 CNN network structure

The convolutional layer uses the feature picture matrix of l convolution cores of the same size and the fused badminton movement to generate a new pixel matrix of l channels.

Sub-sampling layer: The most important part of the sub-sampling layer is pooling, also called the pooling layer (Jie et al. 2020). The Pooling layer is divided into average pooling and maximum pooling. Average pooling refers to summing all the values in the pooled domain and taking its average as the value in the sub-sampling feature map during the pooling process; The maximum pooling takes the maximum value in the pooling domain as the characteristic value of sub-sampling.

Set the input badminton take-off action characteristic graph matrix F, the subsampling pool domain is matrix P of $c \times c$ the offset is b_2, and the characteristic diagram of the sub-sampled badminton take-off action is as S, set the moving step size of the pool process as c. The algorithm expressions of average pooling and maximum pooling are respectively:

$$\overline{S}_{ij} = \frac{Relu\,(x)}{c^2}\left(\sum_{i=1}^{c}\sum_{j=1}^{c} F_{ij}\right) + b_2 \tag{13}$$

$$S_{ij-\max} = \mathrm{Relu}\,(x)\underset{i=1,j=1}{\overset{c}{\max}}\,(F_{ij}) + b_2 \tag{14}$$

In the formula, $c_{max}^{i=1,j=1}$ represents the largest element removed from the pooling domain of the size of the input feature figure F as $c \times c$.

Full connection layer: In the full connection layer, the Softmax classifier is used to classify the characteristics of badminton take-off after convolution and pooling and judge the probability that the badminton movement jump action characteristics x belongs to j, Soft-Max regression model is an extension of the Logistic regression model, which is mainly used to solve multi-classification problems. The classification results of the Softmax classifier are as follows:

$$s^{(i)} = \frac{S_{ij-\max} - \bar{S}_{ij}}{\sum_{l=1}^{k} e^{\theta_{\tau j} n_{epo} x^{(i)}}} \times e^{\theta_{\tau j} n_{epo} x^{(i)}} \tag{15}$$

In the formula, k is the total number of layers, θ_τ is the parameter variable of the model, n_{epo} is the number of iterations.

Output layer: The output layer outputs the recognition results of badminton take-off actions classified by softmax.

1.5.2 Improvement of pool layer based on adaptive pool

First, the pool layer is improved to optimize the feature extraction process of badminton take-off.

In the convolutional neural network learning process, there will be many different feature maps and pools, and it isn't easy to achieve a satisfactory result when facing these feature maps and pools. A dynamic adaptive pooling method based on a maximum pooling algorithm is proposed to improve the pooling layer further. This model can dynamically adjust the pooling process according to different characteristic diagrams of badminton take-off, and adaptively adjust the pooling weight according to the content of each pooling domain. If the pool domain has only one value, it is both the maximum value and the representation of its characteristics. If the eigenvalues of this pooled domain are all the same, its maximum value can also be expressed as the eigenvalues of the pooled domain. Therefore, based on the maximum pooling algorithm, a mathematical model is constructed to simulate the function according to the interpolation principle. Set μ is the pooling factor, then the expression of the dynamic adaptive pooling algorithm is:

$$V_{ij} = \mu \underset{i=1,j=1}{\overset{c}{\max}} \frac{(F_{ij})}{s^{(i)}} + b_2 \tag{16}$$

This formula is the basic expression of a dynamic adaptive algorithm. Its essence is to optimize the maximum pooling algorithm using the pooling factor μ. The optimized characteristics of badminton take-off action can express badminton take-off action more accurately and help realize accurate badminton take-off action recognition. The other parameters follow the parameter setting of the maximum pool layer.

$$\mu = \rho \frac{a\left(v_{\max} - a\right)}{v_{\max}^2 V_{ij}} + \theta \tag{17}$$

In the formula, a is the average value of the pooled domain elements except the maximum value, v_{max} is the maximum value among pooled domain elements, θ is to correct the error term ρ is the characteristic coefficient, and the calculation expression is:

$$\rho = \frac{c}{1 + \left(n_{epo} - 1\right) c^{n_{epo}^2 + 1}} \tag{18}$$

Pool factor $\mu \in (0, 1)$, in this way, both maximum pooling and average pooling can be considered. The accuracy will not be lost when dealing with the pool domain with obvious maximum characteristics, and the influence of maximum pooling can be weakened when dealing with other pool domains so that convolutional neural networks can extract more accurate features when dealing with different pool domains under different iterations, and the recognition effect of badminton take-off action can be improved.

1.5.3 Optimization of recognition results based on batch normalization

When using CNN to identify badminton take-off action, the essence of CNN's learning process is to learn the data distribution of badminton take-off action feature fusion results. The training of the network is a complex process. As long as the first few layers of the network change slightly, the later layers will be accumulated and enlarged. Once the distribution of input data in a certain layer of the network changes, then this layer of the network needs to learn this new data distribution. During the training process, the distribution of each layer of training data has been changing, and the learning rate required by each layer is different. Usually, the minimum learning rate is needed to ensure the effective decline of the loss function, which will affect the training speed of the network. Therefore, the batch-normalization (BN) algorithm is added to solve the problem of changing data distribution in the middle layer during training.

The BN algorithm normalizes the data of each layer to a mean value of 0 and a standard deviation of 1, which makes the data stable. Therefore, it can use a large learning rate to train, accelerate the convergence of the network and improve the training speed. It also improves the efficiency of badminton take-off recognition. BN algorithm mainly uses the following formula for normalization:

$$\widehat{x}^{(k)} = \frac{x^{(k)} - E\left[x^{(k)}\right]}{\mu\sqrt{Var\left[x^{(k)}\right]}} \tag{19}$$

Among them, $E\left[x^{(k)}\right]$ refers to the mean value of each batch of training data x^k; The denominator is one standard deviation of x^k of each data batch. In the following processing, Formula (19) will also be used to normalize the input data of a certain layer network. It should be noted that the batch random gradient descent method is used in the training process.

If only Formula (19) is used to naturalize the output data of a certain layer of the network and then send it to the next layer, it will affect the learned characteristics of this network layer. So, the key of the BN algorithm is to introduce parameters $_{\gamma}$ and β; the formula is:

$$y^{(k)} = y^{(k)}\widehat{x}^{(k)} + \beta^{(k)} \tag{20}$$

Among them $\widehat{x}^{(k)}$ represents data normalized to a mean of 0 and a standard deviation of 1; The function of γ and β is to maintain the expressive ability of CNN model during training. For example, in the middle part of the sigmoid activation function, the function is similar to a linear function. After using BN, the normalized data will only be distributed in the linear part, which will make the generalization ability of the model worse, introduce γ and β will make the data not only distributed in the linear part but also in the nonlinear part, which increases the generalization ability of the model and improves the recognition effect of badminton take-off action. When $y^{(k)} = \sqrt{Var\left[x^{(k)}\right]}$, $\beta^{(k)} = E\left[x^{(k)}\right]$, can completely restore the original data of a certain layer. γ and β are necessary to use the backpropagation algorithm for training. The following formula is the data normalization process of BN:

$$\begin{cases} \frac{1}{m}\sum_{i=1}^{m} x_i \to \phi_B \\ \frac{1}{m}\sum_{i=1}^{m}\left(x_i - \phi_B\right)^2 \to \sigma_B^2 \\ \frac{x_i - \phi_B}{\sqrt{\sigma_B^2 + \varepsilon}} \to \widehat{x}_i \\ BN_{\gamma,\beta}(x_i) \equiv \gamma\widehat{x}_i + \beta \to y_i \end{cases} \tag{21}$$

Among them x_i is the input value, m is that numb of batch input data, ϕ_B and σ_B^2 are mean and variance, respectively, $y_i = BN_{\gamma,\beta}(x_i)$ is the training output result of batch normalization, that is, the final result of badminton take-off recognition.

Table 1 Computer parameters

Hardware	Argument
CPU	I7-6700
Graphics card	GTX—1030
Internal memory	16 GB
Operating system	windows10
Hard disk	1 T

After the optimization of batch normalized CNN training, the recognition efficiency of badminton take-off can be significantly improved.

2 Experimental analyses

2.1 Experimental setup

To verify the application effect of this method, the video of five badminton games of five badminton players is used as the experimental sample data, and the take-off action of badminton is recognized by this method. The computer hardware parameters related to identification are set, as shown in Table 1.

To verify the tracking effect of the athletes' movements in the badminton sports video of this method, taking the competition video of a badminton player as an example, the athletes' movements are tracked by this method, and some tracking results are shown in Fig. 8.

From the analysis of Fig. 8, it can be seen that the method in this paper can accurately track the movements of badminton players in the video, in which the red rectangular area is the tracked badminton player's target. The tracked badminton player's movement tracking image provides a good data basis for the static feature extraction of the subsequent badminton take-off movement.

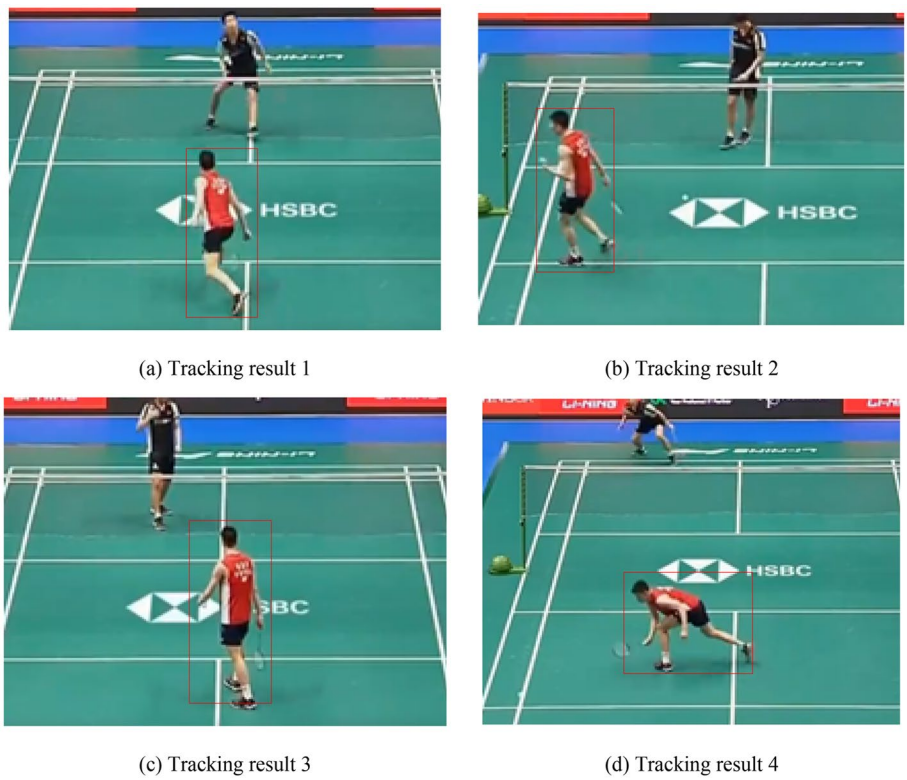
2.2 Analysis of results

To verify the actual recognition results of this method, one of the game videos is selected for recognition in the human skeleton sequence of the video is shown in Fig. 9.

It can be seen from Fig. 9 that the human skeleton feature model collected by this method is very accurate, which lays a foundation for the subsequent collection of dynamic features.

To ensure that the improved convolutional neural network can successfully identify the take-off action, it is necessary to train the network, and the training results are shown in Fig. 10.

As seen in Fig. 10, the network's recognition effect is improving with increased training times. The loss rate decreases rapidly in the first 20 training times, reaches the optimal value in the 50 training times and gradually converges after 50–100 training times. After 500 training times, the loss rate of the network is about 0.25%. The accuracy is different. After only about 20 trainings, the accuracy



(a) Tracking result 1

(b) Tracking result 2

(c) Tracking result 3

(d) Tracking result 4

Fig. 8 Athlete action tracking results in badminton sports video

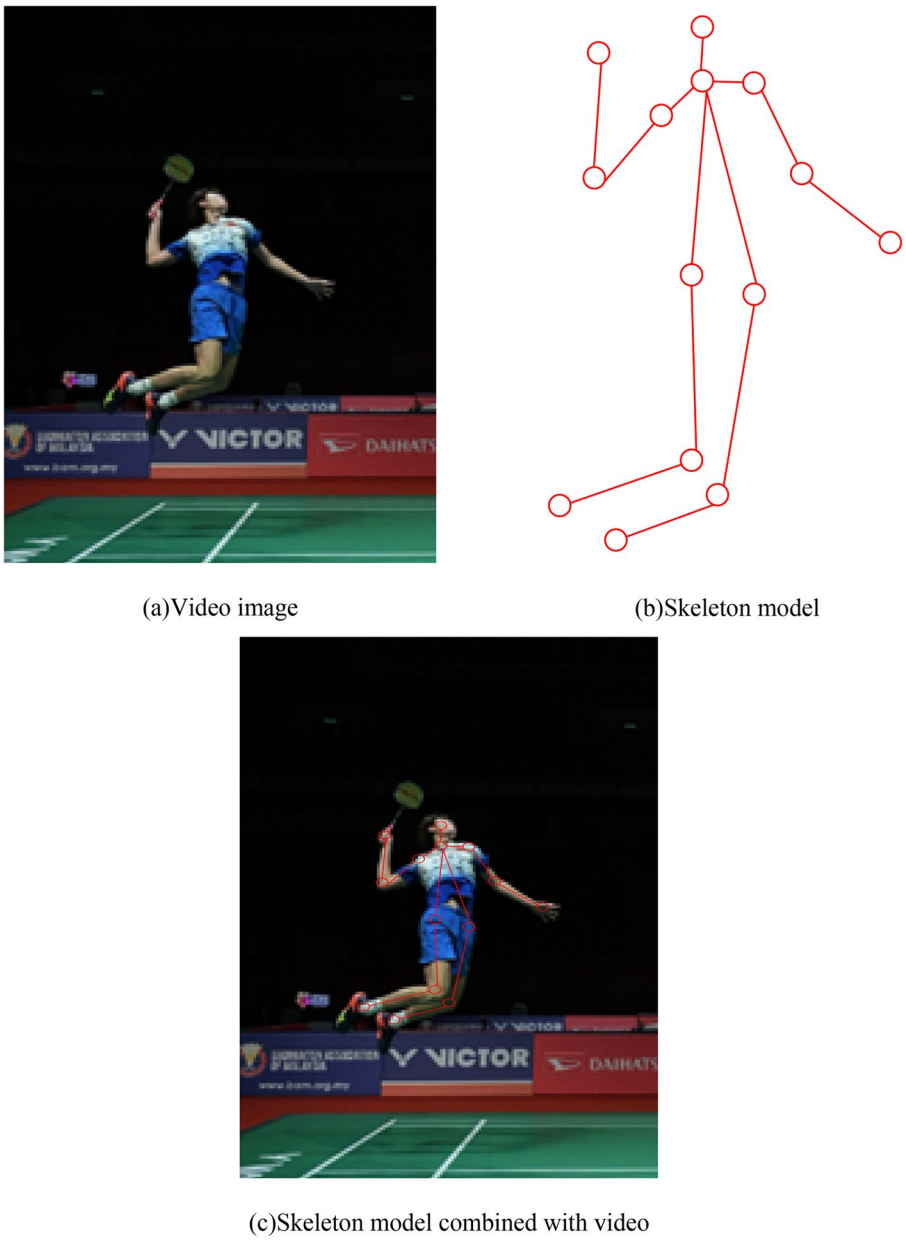


Fig. 9 Human skeleton model in the video

(a)Video image (b)Skeleton model

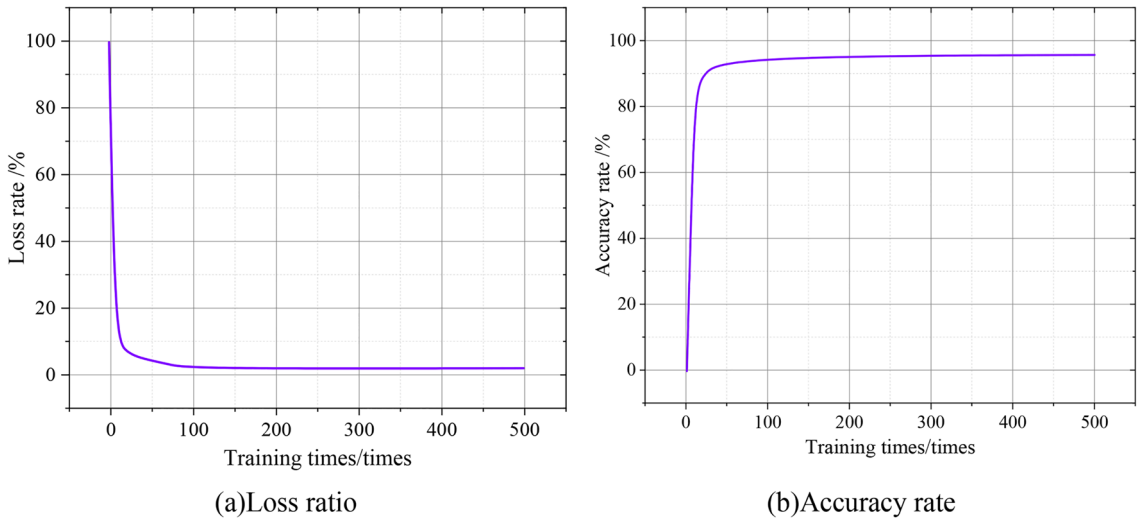
(c)Skeleton model combined with video

reaches nearly 90%, and after 50 trainings, the results begin to converge, and after 500 trainings, the accuracy reaches about 98%. It can be seen that this method has a very high accuracy after training.

The convolution layer of the convolutional neural network is mainly responsible for the feature extraction of badminton take-off action, and the number of convolution layers directly affects the training ability of the network. Under the number of 1–13 convolution layers, the training samples of the CNN neural network are trained by this method, and the training error of the target network is set to 0.001. The training results of different numbers of convolution layers are shown in Table 2.

Analysis of Table 2 shows that when the number of convolution layers of the CNN network is 9, the training times of the CNN network are 120 times, and the training error of network training is 0.000997, which is the closest to the set network training target. Therefore, when the number of convolution layers of the CNN network is 9, the optimal network structure is constructed, improving the network training ability and benefiting the feature extraction of badminton take-off action.

To verify the multi-modal feature recognition method used in this paper, the improved CNN method is used to test the dynamic, static, and fusion features after training. The recognition results of the three features are shown in Fig. 11.



(a)Loss ratio (b)Accuracy rate

Fig. 10 Network training results

Table 2 Training results of different quantity of convolutional layers

Number of convolutional layers/pieces	Training frequency	Network training error
1	278	0.000943
2	257	0.000941
3	235	0.000944
4	208	0.000943
5	178	0.000986
6	162	0.000977
7	138	0.000902
8	124	0.000949
9	115	0.000997
10	120	0.000994
11	122	0.000938
12	124	0.000938
13	121	0.000992

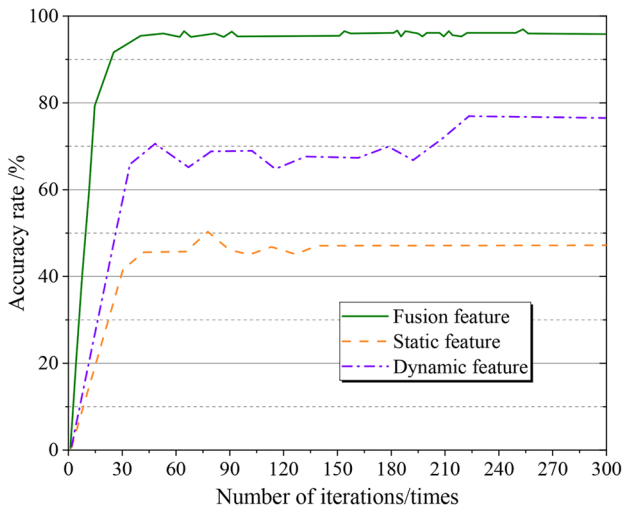


Fig. 11 Recognition results of different features

As can be seen from Fig. 11, the recognition accuracy of only using static features is less than 50%, which is far from the use requirement, while the recognition accuracy of dynamic features is only 80% at the highest. After the fusion of the two features, the feature recognition rate can reach about 97%, so it is the best choice to choose fused features as the input to improve CNN.

To verify the recognition accuracy of the improved CNN network, badminton players' competition videos are selected for recognition. It is combined with traditional CNN, improved motion recognition based on separable spatio-temporal attention of bone and video preprocessing (method in Reference (Climent-Perez et al. 2021)), motion and interaction recognition based on the multi-view representation of manual low-level skeleton features (method in reference (Avola et al. 2022)), motion recognition from 3D bone data using body state (method in reference (Mokari et al. 2017)), video motion recognition based on pseudo-3D residual attention network (method in reference (Chen et al. 2022)), multimodal human behavior recognition in autonomous systems based on environmental intelligence (reference (Jain et al. 2023) method) to make a contract, and the results are shown in Fig. 12.

As shown in Fig. 12, among the above methods, the method in Reference (Mokari et al. 2017) has the worst performance, and the accuracy is less than that of the traditional CNN method in 300 iterations. In the first 150 iterations, the recognition accuracy of the method in reference (Jain et al. 2023) is less than that of the method in reference (Mokari et al. 2017); after 150 iterations, the accuracy exceeds that of the method in reference (Mokari et al. 2017); after 240 iterations, the accuracy exceeds that of the classical CNN method, and after 300 iterations, the accuracy is about 70%, ranking fourth among the above methods. The accuracy

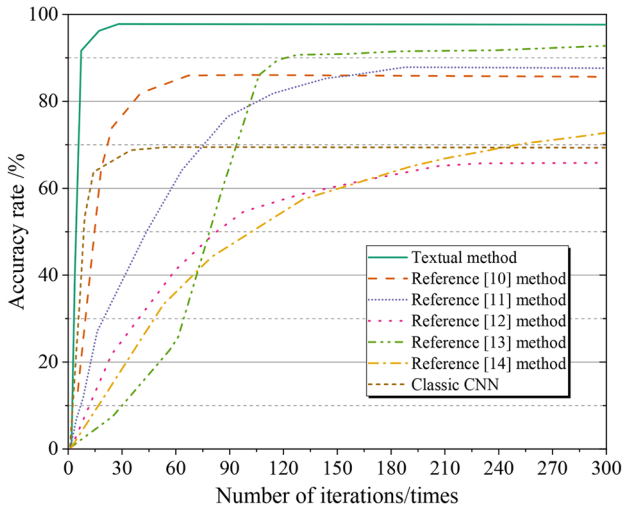


Fig. 12 Monitoring results of different methods

Table 3 Comprehensive results of identification of takeoff actions in badminton sports

Category	Recall	Precision	Accuracy	F1
Proposed method	0.989	0.991	0.978	0.988
Traditional CNN method	0.754	0.821	0.796	0.801
Reference (Climent-Perez et al. 2021) Method	0.846	0.852	795	0.864
Reference (Avola et al. 2022) Method	0.911	0.894	0.915	0.884
Reference (Mokari et al. 2017) Method	0.852	0.945	0.923	0.842
Reference (Chen et al. 2022) Method	0.852	0.798	0.917	0.854
Reference (Jain et al. 2023) Method	0.912	0.852	0.914	0.924

of the methods in references (Climent-Perez et al. 2021), (Avola et al. 2022) and (Chen et al. 2022) exceeded 80% in 300 iterations, and the accuracy of reference (Chen et al. 2022) was about 90%, but the convergence time of reference (Chen et al. 2022) was long, and the accuracy converged after 120 iterations. Among the three methods, the convergence speed of reference (Climent-Perez et al. 2021) began to converge after 60 iterations. The method presented by the current research is superior to other methods in many aspects, and its convergence speed is fast. After 30 iterations, it converges to the optimal result with high accuracy. After 300 iterations, the accuracy is about 97%. To sum up, it can be proved that the method in this paper can be very practical.

Introduce Recall, Precision, Accuracy, and F_1 index is the evaluation index of badminton take-off recognition by the method in this paper, and the correct rate is the ratio of the samples that identify badminton take-off with the actual action to the total sample. The accuracy rate is that the badminton take-off action consistent with the actual action

Table 4 Identification results

ID	Time	Identification result	Actual result	Jump or not
1	1:10	Run	Run	N
2	2:15	Bat	Bat	N
3	2:40	Smash	Smash	Y
4	3:21	Run	Run	N
5	5:33	Bat	Bat	N
6	6:17	Tumble	Tumble	N
7	7:50	Smash	Smash	Y
8	10:25	Smash	Smash	Y
9	15:53	Bat	Bat	N
10	17:29	Bat	Bat	N

accounts for the proportion of all samples in identifying the badminton take-off action. The recall rate is to identify the take-off action of badminton and the actual action, all of which is the ratio of the take-off action of badminton to all action samples; the F index is the balanced average of recall rate and accuracy rate. The greater the value of the above indicators, the stronger the comprehensive performance of the method for identifying the take-off action of badminton. Taking a badminton player's competition video as an example, this paper uses the traditional CNN method and five literature methods to identify the take-off action of badminton and counts the Recall, Precision, Accuracy and F_1 index results are represented in Table 3.

It is not difficult to see from the analysis of Table 3 that the evaluation index value of badminton take-off recognition under this method is the maximum value of all methods, which is significantly higher than the literature comparison method and the traditional CNN method, among which the recall rate, accuracy rate, accuracy rate and F_1 index values are 98.9%, 99.10%, 97.80% and 98.80% respectively, which shows that this method has the best performance for badminton take-off recognition.

Choose a game video and use this method to identify the take-off action of badminton. In this game, 10 shots are randomly selected, and the recognition results are shown in Table 4.

As seen from Table 4, in a badminton match, there are many actions, including running, hitting, spiking, and even falling. This method can accurately identify the above actions and can distinguish whether to take off when doing the above actions. By randomly collecting 10 movements, it is evident that the method in this paper is very accurate and practical.

The recognition results of the badminton take-off at 2:40 and 10:25 are shown in Fig. 13.

As can be seen from Fig. 13, this method completes the accurate identification of badminton take-off action, which can provide a basis for the later game resumption and the planning guidance of take-off action.

Fig. 13 Takeoff action recognition results

(a)2: 40 Identification results (b)10:25 Identification results

3 Conclusion

Experiments indicate that the human joint model established by the badminton take-off recognition method based on improved deep learning can accurately display the human body features in the video, and the recognition accuracy of multi-modal features is much higher than that of single-modal features. The recognition accuracy of the improved CNN network is much higher than that of other recognition methods before and after the improvement, and the actions of players in the competition can be accurately classified and whether they are take-off actions can be identified.

This new method based on improved deep learning combines static and dynamic feature extraction to achieve accurate recognition of badminton take-off action through a deep learning network. The method uses 3D ConvNets and LSTM networks to extract features from different angles and improves accuracy through fusion and optimization. To successfully integrate this method into practical applications, it is necessary to consider the complexity of the actual scene, the cost of data collection and model training, real-time requirements, and other factors. Further engineering implementation and system optimization will be key to applying this method to real-world scenarios, and overall, this method has the potential to be applied in areas such as badminton competitions. In other words, this new method of badminton take-off recognition based on improved deep learning is easy to apply to the actual badminton strategy, mainly for the following reasons. Firstly, it enhances accuracy by combining static and dynamic feature extraction through a deep learning network. This enables players to more precisely identify their opponents' take-off moments, allowing them to formulate more accurate coping strategies. Secondly, real-time optimization is achieved by optimizing the CNN network structure and incorporating adaptive pooling and other technologies. This improves the model's response speed, enabling timely capture of the opponent's take-off actions during live games. Consequently, it provides support for the formulation of real-time strategies. Thirdly, the utilization of a human skeleton model assists in extracting action features and facilitates the analysis and comprehension of athletes' movements. This method (using Data-assisted analysis) aids in formulating personalized tactics and strategies. Lastly, the method benefits from the advancements in deep learning technology and the availability of comprehensive open-source libraries and tools. As a result, it can be easily implemented in the engineering aspect of actual badminton strategies, offering standardized operability.

Author contribution Zhang Haiying: Writing—original draft preparation, conceptualization, supervision, project administration. Lu Lianju: methodology, software, validation

Availability of data and materials On request.

Declarations

Competing interests The authors declare no competing of interests.

References

Arashpour M, Kamat V, Heidarpour A, Hosseini MR, Gill P (2022) Computer vision for anatomical analysis of equipment in civil infrastructure projects: theorizing the development of regression-based deep neural networks. Autom Constr 137:104193

Avola D, Cascio M, Cinque L, Fagioli A, Foresti GL (2022) Affective action and interaction recognition by multi-view representation learning from handcrafted low-level skeleton features. Int J Neural Syst 32(10):2250040

Chen B, Tang H, Zhang Z, Tong G, Li B (2022) Video-based action recognition using spurious-3D residual attention networks. IET Image Proc 16(11):3097–3111

Climent-Perez P, Florez-Revuelta F (2021) Improved action recognition with separable spatio-temporal attention using alternative skeletal and video pre-processing. Sensors 21(3):1005

Davtalab O, Kazemian A, Yuan X, Khoshnevis B (2022) Automated inspection in robotic additive manufacturing using deep learning for layer deformation detection. J Intell Manuf 33(3):771–784

Harvat M, Martín-Guerrero JD (2022) Memory degradation induced by attention in recurrent neural architectures. Neurocomputing 502:161–176

Jain V, Gupta G, Gupta M, Sharma DK, Ghosh U (2023) Ambient intelligence-based multimodal human action recognition for autonomous systems. ISA Trans 132:94–108

Jie HJ, Wanda P (2020) RunPool: a dynamic pooling layer for convolution neural network. Int J Comput Intell Syst 13(1):66–76

Jurado JM, Padrón EJ, Jiménez JR, Ortega L (2022) An out-of-core method for GPU image mapping on large 3D scenarios of the real world. Futur Gener Comput Syst 134:66–77

Kesavavarthini T, Rajesh AN, Venkata Srinivas C, Kumar TVL (2023) Bias correction of CMIP6 simulations of precipitation over Indian monsoon core region using deep learning algorithms. Int J Climatol. https://doi.org/10.1002/joc.8056

Khaddam HS, Ahmad GG (2022) A method to evaluate the diameter of carded cotton yarn using image processing and artificial neural networks. J Textile Inst 113(8):1648–1657

Kumar, C., Subramaniam, G., & Jasper, J. (2022). A novel ROA optimized Bi-LSTM based MPPT controller for grid connected hybrid solar-wind system. COMPEL-The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, ahead-of-print.

Leibovich M, Papanicolaou G, Tsogka C (2020) Synthetic aperture imaging and motion estimation using tensor methods. SIAM J Imag Sci 13(4):2213–2249

Mokari, M., Mohammadzade, H., & Ghojogh, B. (2017). Recognizing involuntary actions from 3D skeleton data using body states. ArXiv Preprint.

Ozdemir N, Esen H, Secer A, Bayram M, Yusuf A, Sulaiman TA (2021) Optical soliton solutions to Chen Lee Liu model by the modified extended tanh expansion scheme. Optik 245:167643

Peng W, Shi J, Varanka T, Zhao G (2021) Rethinking the ST-GCNs for 3D skeleton-based human action recognition. Neurocomputing 454:45–53

PlÖtz T (2021) Applying machine learning for sensor data analysis in interactive systems: common pitfalls of pragmatic use and ways to avoid them. ACM Computing Surv (CSUR) 54(6):1–25

Rico VJ (2022) Long sigmoid and twisted *ascospores* in the genus *Harpidium*: H *longisporum* sp nov a synopsis of the genus and a key to the species. Lichenol 54(3–4):175–181

Shin D (2023) Embodying algorithms, enactive artificial intelligence and the extended cognition: you can see as much as you know about algorithm. J Inf Sci 49(1):18–31

Sun Y, Xu J, Lin G, Ji W, Wang L (2020) RBF neural network-based supervisor control for maglev vehicles on an elastic track with network time delay. IEEE Trans Industr Inf 18(1):509–519

Sun Y, Xu J, Wu H, Lin G, Mumtaz S (2021) Deep learning based semi-supervised control for vertical security of maglev vehicle with guaranteed bounded airgap. IEEE Trans Intell Transp Syst 22(7):4431–4442

Tang W, Yang Q, Hu X, Yan W (2022) Deep learning-based linear defects detection system for large-scale photovoltaic plants based on an edge-cloud computing infrastructure. Sol Energy 231:527–535

Thangarajan SK, Chokkalingam A (2021) Integration of optimized neural network and convolutional neural network for automated brain tumor detection. Sens Rev 41(1):16–34

Totaro S, Hussain A, Scardapane S (2020) A non-parametric softmax for improving neural attention in time-series forecasting. Neurocomputing 381:177–185

WEN ZHH, Zhou M (2020) Recognition of blowholes and cracks on surface of magnetic tile based on deep learning. Ordn Mater Sci Eng 43(6):106–112

Ying Z, Li M, Yan Z, Haiyong C (2020) Application of improved CNN in solar panel defect detection. Computer Simul 37(3):458–463