Research Article

Research on Sports Training Action Recognition Based on Deep Learning

Peng Wang

College of Physical Education, Zhengzhou University, Zhengzhou 450000, China

Correspondence should be addressed to Peng Wang; lyt@zznu.edu.cn

With the rapid development of science and technology in today's society, various industries are pursuing information digitization and intelligence, and pattern recognition and computer vision are also constantly carrying out technological innovation. Computer vision is to let computers, cameras, and other machines receive information like human beings, analyze and process their semantic information, and make coping strategies. As an important research direction in the field of computer vision, human motion recognition has new solutions with the gradual rise of deep learning. Human motion recognition technology has a high market value, and it has broad application prospects in the fields of intelligent monitoring, motion analysis, human-computer interaction, and medical monitoring. This paper mainly studies the recognition of sports training action based on deep learning algorithm. Experimental work has been carried out in order to show the validity of the proposed research.

1. Introduction

In recent years, human motion recognition has become a hot issue in the field of the application system and academic research. As early as 1973, a psychologist named Johansson carried out the motion perception experiment of moving light spot, which is the first modern research on human motion recognition. Since then, until the 1990s, people began to pay more attention to this field. So far, many researchers around the world have done a lot of research on human motion recognition technology. The traditional research on human motion recognition can be divided into the following parts: representation of motion information and recognition and classification of motion information.

Computer vision is the field of artificial intelligence which is mostly used for creating systems to prepare computer for understanding and eliminating issue with artificial images and sense [1]. Human action recognition is having subtask of collective activity recognition for which the available datasets are commonly inadequate. The study has been presented to look into the issues presenting the collective sports dataset containing multitask recognition for sports and collective activity categories.. A novel protocol of

evaluation called unseen sports is presented in which the training and test are carried out on disjoint sets of sports categories [2]. The study proposed human action recognition through deep multimodal feature fusion algorithm [3]. The research fuses visual feature, probability maps, skeleton, and audio signal into hybrid feature utilized for representing human action. Categories of human and nonhuman are classified through the use of convolutional neural network [4]. Research has been done for action recognition of swimming sports based on wireless sensor and field programmable gate array [5].

In recent ten years, deep learning has become a research hotspot in the field of artificial intelligence, and various research results based on deep learning methods have been applied to practice. The sudden boom of deep learning is not accidental, but the reward of decades of intensive work of researchers in this field. From the 1940s to 1960s [6–9], the rudiment of deep learning began to appear in cybernetics. In 1958, Rosenblatt designed neuron perceptron and relized the training of a single neuron. In the 1990s, the emergence of the backpropagation algorithm made it possible to train neural networks with one or two hidden neurons. It was not until 2006 that the concept of deep learning was formally

established by Hinton et al., which set off the third wave of deep learning.

Following are the main contributions of the study:

(i) To study the existing approaches in the context of sports training recognition

(ii) To study the recognition of sports training action based on deep learning algorithm

(iii) Experimental work has been carried out in order to show the validity of the proposed research

2. Human Motion Video Image and Motion Information Representation

2.1. Motion History Image. The motion history image was first proposed by Davis and Bobrick. Before that, they first proposed a binary motion energy image, which is the predecessor of the motion history image. So, let us take a look at the motion energy image. The motion energy image mainly describes how the object moves and space changes, to recognize the moving object. It can describe the outline of the object movement and the spatial distribution of the energy [10–12].

As shown in Figure 1, we take the action of sitting down as an example. The upper line is the keyframe of the action, and the next line shows the binary motion image accumulated from the start frame to the corresponding frame. We can observe that the blank area in the image is the target motion area. By observing the shape of the moving area of the target, the occurrence of the movement and the observation angle is judged [13–15].

We call the accumulated binary motion image the motion energy graph, as shown in the following equation:

$$E_\tau = (x, y, t) = \sum_{i=0}^{\tau-1} D(x, y, t - i), \tag{1}$$

where $E_\tau = (x, y, t)$ is the binary motion energy image, $D = (x, y, t)$ is the frame difference between the t frame and the $t - 1$ frame, and the motion energy image $E = (x, y, t)$ is the cumulative sum of the frame differences.

Although a motion energy map can reflect the spatial information of motion, it cannot reflect its temporal information. Therefore, the motion image emerges as the times requirement based on the motion energy image. By calculating the pixel changes in the same position at a certain time, it presents the target motion in the form of image brightness. This method belongs to the template method based on vision. The gray value of each pixel in the motion history image shows the motion of position pixels in the video sequence. If the last moving time of the pixel is closer to the current frame, the higher the gray value. Compared with the motion energy image, it can not only show the sequence of action but also contain more details. Therefore, the motion history image can represent the movement of the human body in a movement process, which makes it widely used in the field of motion recognition. Let s be the intensity value of pixels in the motion history map, and $S_\tau = (x, y, t)$ is the update function.

$$S_\tau(x, y, t) = \begin{cases} \tau, & \text{if } \Psi(x, y, t) = 1, \\ \max(0, S_\tau(x, y, t - 1) - \delta), & \text{otherwise}, \end{cases} \tag{2}$$

where (x, y) represents the position of the pixel and t is the time; t is the duration, which determines the time range of motion from the angle of frame number; 8 is the attenuation parameter. The update function $S(x, y, t)$ can be defined by optical flow, interframe difference, or image difference, and the interframe difference method is the most commonly used. Its application is shown in formulas (3) and (4):

$$\Psi(x, y, t) = \begin{cases} 1, & \text{if } D(x, y, t) \geq \xi, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where

$$D = (x, y, t) = |I(x, y, t) - I(x, y, t \pm \Delta)|, \tag{4}$$

where $I(x, y, t)$ is the intensity value of the coordinate (x, y) pixel in the T frame of the video image sequence, Δ is the interframe distance, and ξ is a difference threshold given by a human, which can be adjusted with the change of video scene.

Figure 2 shows the effect pictures of motion history images corresponding to different T values. It can be seen from Figures 2(a) and 2(b) that when the value of T is too small, the whole motion trajectory of the action cannot be obtained completely. As shown in Figure 2(d), when the value of R is too large, the change of the intensity value of the motion track in the captured motion history map is not obvious, which leads to the loss of the information of the action time dimension. We cannot distinguish that the value of t must be considered in the motion history map obtained because the value is too small. As for the difference threshold, if the value is too small, the acquired motion history map will exhibit a lot of messy noise. As shown in Figure 2(e), the obtained image cannot distinguish the foreground from the background well; if the value is too large, the area with a smaller pixel intensity value will disappear, and empty holes will appear, resulting in loss of action information. With the increase of the value, the void area will be larger and larger, until the final motion history image only contains the contour edge. Through the experiment, the optimal value is $t = 50$, $\xi = 40$, which can obtain the most sufficient and effective motion trajectory information.

2.2. Rainbow Coding. The pseudocolor processing of the image can transform the image information into a form that is easier to recognize by humans or machines and enhance the useful information in the image. Pseudocolor processing refers to the technical process of converting a black and white gray image or multiband image into a color tone image. The commonly used pseudocolor coding methods are density segmentation, filtering, and gray level color transformation.

The density segmentation method is mainly used to deal with the image with discontinuous hue, which is the simplest
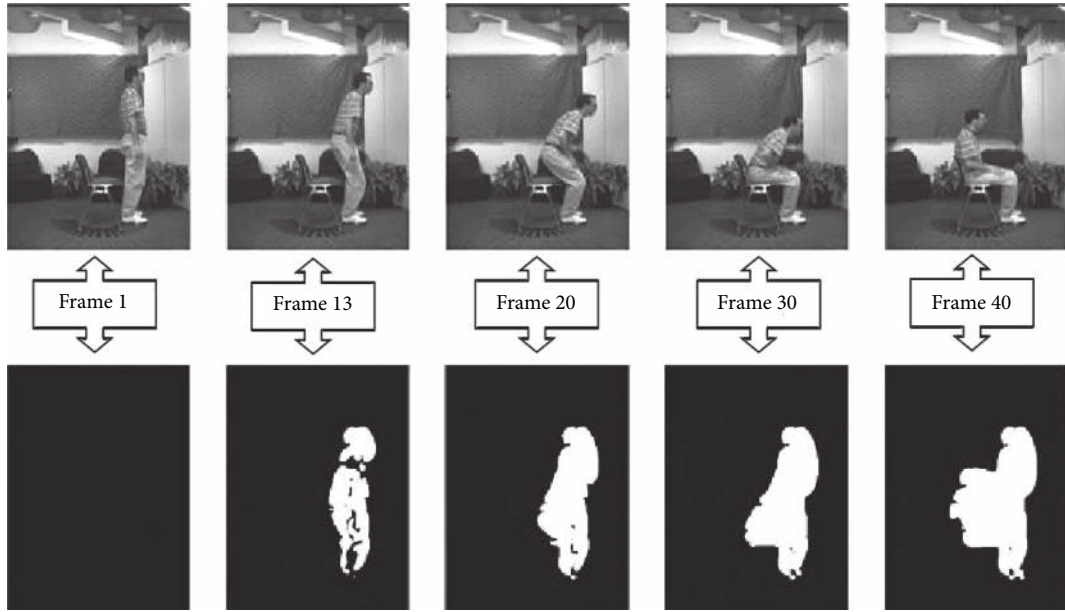
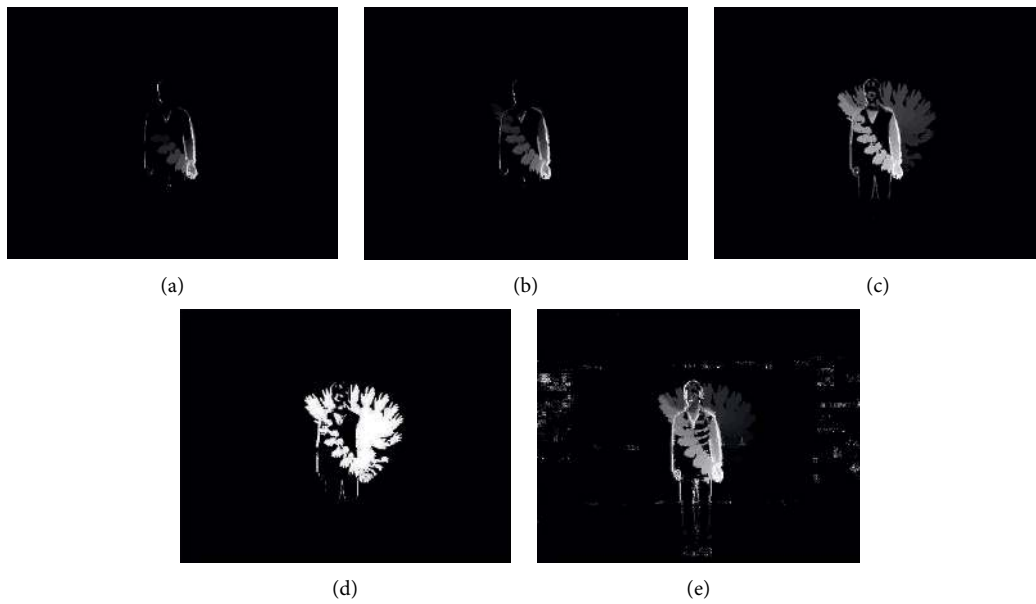FIGURE 1: Keyframe and motion energy diagram of action "sit down."

FIGURE 2: Motion history of different parameters. (a) $\tau = 10$, $\xi = 40$. (b) $\tau = 20$, $\xi = 40$. (c) $\tau = 50$, $\xi = 40$. (d) $\tau = 100$, $\xi = 40$. (e) $\tau = 50$, $\xi = 10$.

method of pseudocolor enhancement. It divides the gray level of a gray image from 0 to 255 into m intervals g, $I = 1, 2, \ldots, M$, and then assigns a specific color C; to each interval, a color image is obtained from a gray image. However, the disadvantage of this method is that the change of hue is not continuous, and the image has obvious blocks, and the number of colors is not rich.

The filtering method is a method based on the frequency domain. It does not rely on the gray level of the image to generate pseudocolor but is determined by the different spatial frequencies of the gray image. As shown in Figure 3,

the gray image is first transformed into the frequency domain by Fourier transform, and then it is separated into three independent variables by using three filters with different characteristics in the frequency domain. Then, three single-channel images with different frequency components are obtained by inverse Fourier transform of these three variables. Then, they are processed, such as histogram equalization. Finally, we synthesize our pseudocolor images as RGB tricolor components.

There are many color transformation methods based on gray levels, such as gray mapping, rainbow coding, and so
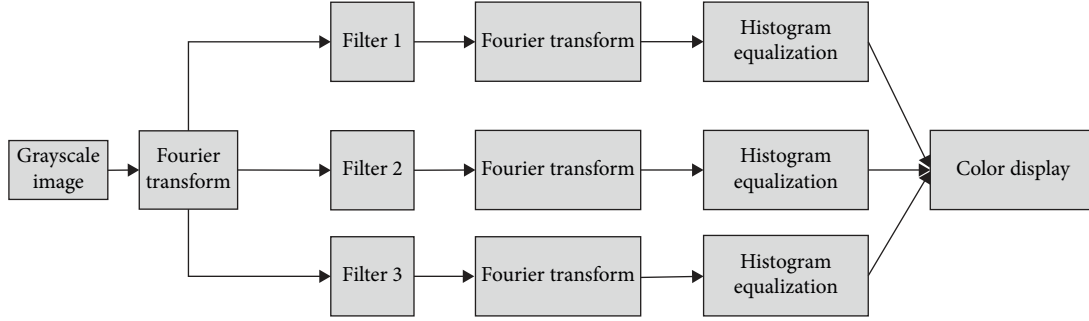
FIGURE 3: Flowchart of the filtering method.

on. But the central idea is based on the principle of color; according to different coding formulas, the gray value of the image is generated into three-channel values of red, green, and blue, and then the color is synthesized. In RGB color space, any color can be composed of red, green, and blue in different proportions. Therefore, what we need to set is the transformation function of the three color channels. The color matching equation is shown in formulas (5)–(7):

$$R(x, y) = T_R\{f(x, y)\}, \tag{5}$$

$$G(x, y) = T_G\{f(x, y)\}, \tag{6}$$

$$B(x, y) = T_B\{f(x, y)\}, \tag{7}$$

where $R(x, y)$, $G(x, y)$, and $B(x, y)$ are the values of red, green, and blue, respectively, $f(x, y)$ is the gray value of (x, y) points on the gray image, and T_R, T_G, and T_B are the corresponding mapping functions. The pseudocolor image we need can be obtained by driving the color display with the three-channel values. It can be seen that the red, green, and blue mapping functions are very important, which determine the quality of pseudocolor after transformation. Different mapping functions will result in different pseudocolor images.

2.3. Improved Motion History Image. It is not effective to extract motion history images from RGB video and send it to the network for training. In this paper, we propose a human motion recognition method based on the improved motion history image, mainly from the following aspects.

2.3.1. Removing Redundant Motion Sequences. In the experiment, it is found that the performer usually has a reaction time of about 1 second when the action execution command is issued. Similarly, after the execution of the action, there is a period of static time, which means that there are useless still frames at the beginning and end of the dataset video. These frames contain useless redundant information and even cover the important information of keyframes, which directly affects the quality of the extracted motion history and then the image. Therefore, before extracting the motion history image from the video, the first

step is to remove 10 frames of each video, and then the motion history image is obtained.

2.3.2. Applying Rainbow Coding. According to Abidi et al. in the report, better perceptual quality and more information can be obtained by encoding gray texture with human perceptible color. Inspired by this, in this paper, we use the rainbow coding to enhance the motion model of the motion image. The larger the gray value, the closer it is to red; otherwise, the smaller the gray value, the closer it is to blue. The motion history image encoded by the rainbow has a rich color. The distribution of color reflects the level of motion and the information of the time dimension, which can more effectively represent the motion information of action.

3. Overview of Deep Learning

3.1. Deep Learning Method. Deep learning is a method of learning data representation in machine learning. Through learning multilevel combination, we can get the recognizable feature representation and finally map the feature representation to the task target. As a kind of machine learning, deep learning is superior to machine learning in that it can automatically learn the feature representation of data. As shown in Figure 4, deep learning avoids the trouble of manual feature design in machine learning; in the traditional machine learning process, data are extracted from input to manual feature extraction, and then the extracted features are mapped to learning objectives. Deep learning simplifies this process. Using the end-to-end idea, the deep learning model can directly convert the input to the output, and the process of feature extraction and feature mapping to the target output is automatically completed by the model, which eliminates many complicated intermediate processes in traditional machine learning.

Like other machine learning methods, the essence of deep learning is to use algorithms to learn knowledge from a large number of data, but it is called "depth." On the one hand, the depth of the deep learning model is the stack of multiple layers of modules, and the number of layers is large. The data from input to target output need multilayer transformation, and the model is deeper; on the other hand, the feature extraction of deep learning is a process of abstraction and fusion from generalization features to semantic features. The shallow features are some basic patterns, the
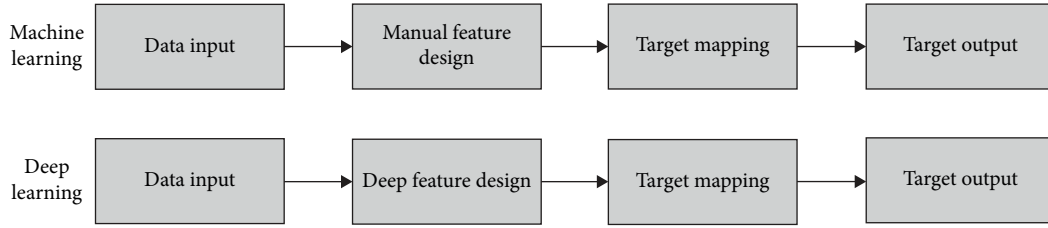
Figure 4: Machine learning method and deep learning method flow.

middle-level features begin to have some fuzzy semantics, and the deep-seated features are recognizable semantic features, which can be mapped to the target output; the deep-seated features have the following features: the obvious progressive process from shallow to deep, and what the model finally learns is this deep feature representation method.

Since the development of deep learning, it includes mathematical analysis, linear algebra, probability theory, mathematical statistics, optimization theory, and numerical calculation. It also includes regularization methods such as random deactivation and batch standardization to ensure the generalization performance of the model, model learning method combining backpropagation and random gradient descent, and distributed representation strategy in the absorption representation learning field. These have injected strong vitality into the deep learning method.

3.1.1. Forward Propagation and Backward Propagation.
In deep learning, feedforward neural networks have a forward propagation process. When the input information passes through a layer of hidden units, after layer by layer conversion, the final output is generated. Such a flow of information is called forward propagation; forward propagation can be regarded as a process of network input processing. When the network is training, the input data flow through the network and produce output and calculate the loss function with the target output. Combined with the backpropagation algorithm, the network is updated; when the network weight parameters are fixed, the network training has been completed, and the forward propagation is a prediction process, and the final prediction results are obtained by the forward propagation of the input.

In contrast to forwarding propagation, backpropagation is a process in which the information of cost function flows forward through the network and calculates the gradient. The nonlinearity of the deep learning model makes the learning of the model nonconvex optimization. Generally, the gradient-based method is used to iterate the training model, so that the cost function of the model converges to the minimum value. The backpropagation algorithm points out a way to calculate the gradient. The gradient points out the optimization direction and then combines the stochastic gradient descent algorithm to update the weight parameters of the model. The core idea of backpropagation is to recursively calculate the gradient of the cost function

concerning hidden layer output and weight using chain rule. Firstly, the gradient of the cost function about the output of the last hidden layer and the gradient of the output of the last layer concerning the weight parameters are calculated, and then the gradient of the cost function concerning the weight parameters of this layer can be obtained by using the chain rule, and then the gradient of the output of this layer concerning the input is calculated and multiplied.

Based on the gradient of the previous cost function on the output of this layer, the derivative of the cost function concerning the output of the penultimate hidden layer is obtained, and so on, until the hidden layer of the lowest layer, to obtain the gradient of the cost function concerning the weight parameters of all layers.

For the deep learning model with L-layer hidden layer, the weight parameter of each layer is w, and the input of each layer is X. To calculate simply, it is assumed that the output of the former layer is the input of the latter layer, X is the output of the last hidden layer, and the cost function is J.

Calculate $(\partial J/\partial X_{L+1})$ and keep it as the next operation. For each layer $l, l = \{L, L-1, \ldots, 1\}$, the calculation process is as follows:

$$\text{calculate } \frac{\partial X_{i+1}}{\partial W_i} \text{ and } \frac{\partial J}{\partial X_{i+1}}, \quad \text{then } \frac{\partial J}{\partial W_i} = \frac{\partial J}{\partial X_{i+1}} \frac{\partial X_{i+1}}{\partial W_i},$$

$$\text{calculate } \frac{\partial X_{i+1}}{\partial X_i}, \quad \text{then } \frac{\partial J}{\partial X_i} = \frac{\partial J}{\partial X_{i+1}} \frac{\partial X_{i+1}}{\partial X_i},$$

$$\text{reserve } \frac{\partial J}{\partial X_i} \text{ for the next operation.}$$

$$(8)$$

In the above operation process, the propagation process is simplified. In practice, in addition to the gradient of the weight parameters, if there are bias terms and regular terms, the gradient of the cost function concerning bias and regularity also needs to be calculated. Moreover, before the hidden layer output, there is generally an activation process, and the gradient of the cost function on the layer output needs to be converted to the gradient before the activation.

The backpropagation algorithm is not only used to calculate the gradient of cost function about parameters but also used to calculate the gradient of other outputs on parameters to analyze the model. The backpropagation algorithm can be used to calculate the gradient of any function, which is a very practical method to calculate the gradient.

The backpropagation algorithm combined with random gradient descent has always been the most commonly used learning method for deep learning model law.

3.1.2. Distributed Representation. When it comes to deep learning, we have to mention distributed representation. As a kind of representation learning, deep learning is unique in that it can automatically learn the distributed feature representation of data according to different learning tasks; as an important tool of representation learning, distributed representation is a representation of concepts expressed by a combination of multiple separated features.

In the deep learning model neural network, neuron to the semantic concept is many-to-many mapping. The semantic concept may be represented by activation patterns distributed in different neurons, and a neuron can participate in the representation of different semantic concepts. For example, the semantic concept "cat" can be represented by the combination of features such as "ear," "four legs," and "fur." In convolutional neural networks, these features are the activation modes of several convolutional neurons, and the feature of "fur" can also be a local feature of the semantic concept "dog," "leopard," or "tiger." The neurons that generate this activation mode can also participate in the representation of these semantic concepts; the advantage of this feature is that fewer learning samples can be used to achieve the same as nondistributed tables.

It shows the same learning effect. For example, for input samples such as "white cat," "black cat," "white dog," and "black dog," when distributed representation is not applied, four separate neurons are needed to learn the concepts of color and category at the same time. After using distributed representation, only two kinds of neurons are needed, one is used to describe categories, one is used to describe colors, and the other is to describe colors. Color neurons can learn color concepts from input samples of "cat" and "dog," instead of using specific neurons to learn from specified samples.

4. Analysis and Recognition of Sports Video in the Process of Sports Training

4.1. Adaptive Threshold Moving Object Separation Based on Particle Filter Prediction. The separation of moving objects in sports training videos can collect and process the moving objects from the dynamic background, which is the basis of sports video analysis. Sports training video sequence separation in the sports target is the athlete in the video; the adaptive threshold moving object separation algorithm based on particle filter prediction is used to enhance the accuracy of moving object acquisition. The specific process is as follows: firstly, the foreground image in the video is separated by the three-frame difference method, and the background is projected into adjacent video frames according to the camera steady motion model, and the background separation map of each frame is obtained. The method of background subtraction is used to further separate moving objects. Because of the similarity between the

foreground image and the background image, to avoid the separation of the moving objects in the video foreground image mistakenly fused into the background image, it is necessary to separate the coordinate range of the foreground image and obtain the frame threshold of the background image which is not in the coordinate range of the foreground image according to the particle filter method, to complete the adaptive threshold separation of the moving object.

The foreground target obtained by the three-frame difference method is interfered with by noise, which will lead to false separation. After filtering, it can be used as the operation standard of adaptive threshold separation. Then, through the prediction scheme of the particle filter, the foreground coordinate interval of other frames is predicted according to the result of frame separation at this time. The offset state between the image pixel and the foreground coordinate interval is added, which is set as the foreground separation probability of the pixel in other frames. Based on this probability, the adaptive separation threshold can be calculated.

(1) The probability of separating background points by the three-frame difference method is as follows:

$$P_{L,j} = \begin{cases} 0.1, & \text{the pixel belongs to the foreground,} \\ 0.5, & \text{the background of the pixel is not clear,} \\ 0.9, & \text{the pixel belongs to the background.} \end{cases}$$
(9)

The probability of pixel separation into the background can be obtained by formula (9). For pixels adjacent to the image boundary, it is not clear that they are the background, so a median value of 0.5 needs to be set. The mean filtering method can restrain the bad invasion of noise. After filtering the collected image pixels with a 3×3 filter, the new background separation probability is obtained as follows:

$$P_f(i, j) = \frac{1}{9} \sum_{m=-1}^{1} \sum_{n=-1}^{1} P(i + m, j + m).$$
(10)

(2) The probability of background points is obtained by particle filter. From the schematic diagram of particle filter prediction described in Figure 5, it can be seen that the particle set composed of weighted particles can be regarded as the foreground range of the moving target, and the prediction of foreground range is completed. The particle types in the particle set are vectors (including the x-coordinate and y-coordinate of the upper left corner of the moving object, the x-coordinate and y-coordinate of the lower right-hand corner of the moving object, and the horizontal and vertical movement speed of the upper left-hand corner and the lower right-hand corner of the moving target). After separating the foreground coordinate interval, the actual weights of various types of particles are calculated according to the foreground coordinate interval, and the particle
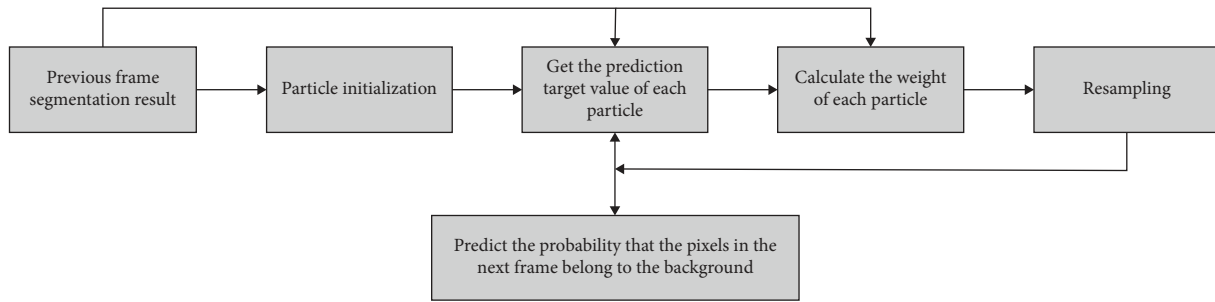
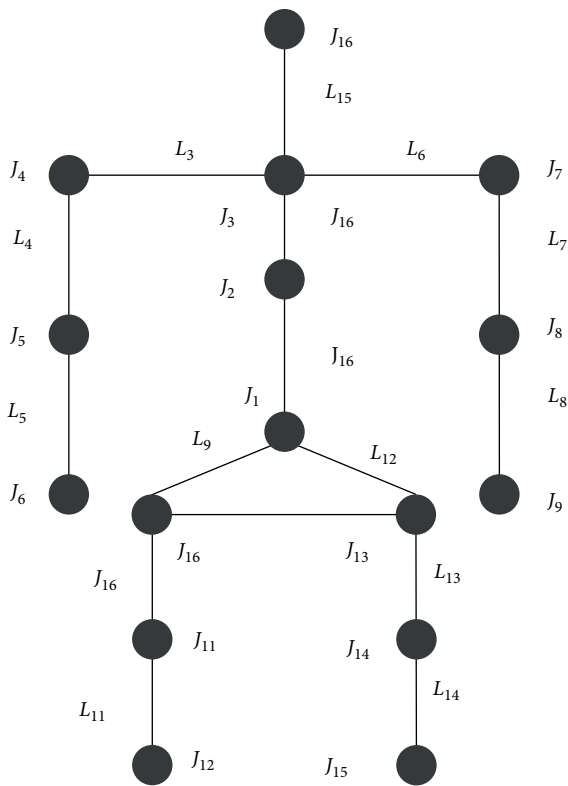FIGURE 5: Schematic diagram of particle filter prediction.

FIGURE 6: Two-dimensional nodes of human skeleton model.

samples are extracted; the particles with higher weight are more likely to be output as samples. The new particle set is formed by sampling the particles, and then the foreground range of subsequent frames is predicted to obtain the probability of different pixels as background in the subsequent frames. If the moving object pixel in the moving video falls in the background probability area, the point belongs to the background; otherwise, it belongs to the moving foreground, and finally, the moving object separation is completed.

4.2. Moving Object Tracking in Moving Video. The purpose of sports target separation in sports training video is to accurately track the sports target, and the purpose of tracking is to collect the motion parameters of athletes' body joints from the sports training video. The adaptive particle filter algorithm is used to track the sports video in the process of sports training, and the human skeleton model is created, as shown in Figure 6.

The human motion model can predict the immediate motion state according to the motion state of the previous frames. In the process of sports training, the sports trend and the number of sports have regularity.

5. Conclusion

The key to the success of human motion recognition is to capture the spatiotemporal motion patterns of all parts of the human body at the same time. In this paper, a deep learning method based on the motion characteristics of local joints is proposed to recognize the motion samples. Then, the effectiveness of the method is evaluated on two datasets. Aiming at the shortcomings of the model, the average error of the model is reduced and the spatial configuration information is introduced to improve the recognition accuracy of the model. The work of this paper is summarized as follows: after the in-depth investigation on the related tasks of motion recognition, the research prospect and application significance of the deep learning method are pointed out, and the recent research status based on the deep learning method is introduced according to the data mode used, which provides the basis for the later work of action recognition based on local motion features.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] X. Xie, "Image recognition of sports training based on open IoT and embedded wearable devices," *Microprocessors and Microsystems*, vol. 82, Article ID 103914, 2021.

[2] C. Zalluhoglu and N. Ikizler-Cinbis, "Collective sports: A multi-task dataset for collective activity recognition," *Image and Vision Computing*, vol. 94, Article ID 103870, 2020.

[3] E. Zhou and H. Zhang, "Human action recognition toward massive-scale sport sceneries based on deep multi-model feature fusion," *Signal Processing: Image Communication*, vol. 84, Article ID 115802, 2020.

[4] W. Yang, Y. Gao, and F. Zhai, "Simulation of sports action picture recognition based on FPGA and convolutional neural network," *Microprocessors and Microsystems*, vol. 80, Article ID 103593, 2021.

[5] T. Liang, "Swimming sports action recognition based on wireless sensor and FPGA," *Microprocessors and Microsystems*, Article ID 103433, 2020, In press.

[6] C. Gong, H. Wang, R. Li et al., "Video moving object trajectory tracking algorithm based on state-dependent detection," *Modern Electronic Technology*, vol. 39, no. 7, pp. 51–56, 2016.

[7] H. Song, F. Wang, X. Liu et al., "Group motion analysis and abnormal behavior detection in a geographical environment," *Geography and Geographic Information Science*, vol. 31, no. 4, pp. 1–5, 2015.

[8] M. Zhu, "Research and application of a three-axis acceleration sensor system in the field of sports technology analysis," *Ice and Snow Sports*, vol. 37, no. 2, pp. 89–96, 2015.

[9] F. Wei, "Loose design of key parts of children's trousers under sports condition," *Xi'an University of Technology Acta Sinica*, vol. 29, no. 5, pp. 550–554, 2015.

[10] H. Guo, "Table tennis decision-making system based on dual-channel target motion detection," *Journal of Guangzhou Institute of Physical Education*, vol. 34, no. 6, pp. 67–70, 2014.

[11] D. Sui and D. Hou, "Research and simulation of dynamic image tracking optimization algorithm," *Modern Electronic Technology*, vol. 39, no. 6, pp. 98–100, 2016.

[12] B. Li, F. Liu, D. Li et al., "Abnormal behavior detection and recognition of moving objects in campus intelligent video surveillance," *Software Guide*, vol. 15, no. 2, pp. 134–136, 2016.

[13] M. Wu and L. Lin, "Research on video recognition and comparison system of competitive sports," *Journal of Guangzhou Institute of Physical Education*, vol. 34, no. 4, pp. 59–61, 2014.

[14] C. Li and X. Lian, "Research on video analysis and intelligent diagnosis system of national traditional sports antagonism," *Journal of Guangzhou Institute of Physical Education*, vol. 36, no. 1, pp. 52–56, 2016.

[15] W. Wei, K. Wu, L. Guo et al., "Video tracking of gradient expansion template motion of memory watershed disc small," *Journal of System Simulation*, vol. 28, no. 2, pp. 462–466, 2016.