Hindawi
Mobile Information Systems
Volume 2021, Article ID 9919992, 10 pages
https://doi.org/10.1155/2021/9919992

Research Article

An Action Recognition Algorithm for Sprinters Using Machine Learning

Fengqing Jiang[1] **and Xiao Chen**[2]

[1]*Institute of Science, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, China*
[2]*Physical Education Department, Zhongnan University of Economics and Law, Wuhan, Hubei 430073, China*

Correspondence should be addressed to Xiao Chen; chenxiaozhongnan@sohu.com

Received 9 March 2021; Revised 14 April 2021; Accepted 5 May 2021; Published 19 May 2021

Academic Editor: Xingwang Li

The advancements in modern science and technology have greatly promoted the progress of sports science. Advanced technological methods have been widely used in sports training, which have not only improved the scientific level of training but also promoted the continuous growth of sports technology and competition results. With the development of sports science and the gradual deepening of sport practices, the use of scientific training methods and monitoring approaches has improved the effect of sports training and athletes' performance. This paper takes sprint as the research problem and constructs the image of sprinter's action recognition based on machine learning. In view of the shortcomings of traditional dual-stream convolutional neural network for processing long-term video information, the time-segmented dual-stream network, based on sparse sampling, is used to better express the characteristics of long-term motion. First, the continuous video frame data is divided into multiple segments, and a short sequence of data containing user actions is formed by randomly sampling each segment of the video frame sequence. Next, it is applied to the dual-stream network for feature extraction. The optical flow image extraction involved in the dual-stream network is implemented by the system using the Lucas–Kanade algorithm. The system in this paper has been tested in actual scenarios, and the results show that the system design meets the expected requirements of the sprinters.

1. Introduction

Scientific and reasonable sports training [1–3] is a sports training mode based on the feedback information of functional monitoring indicators and technical monitoring indicators to regulate the intensity of sports training. At present, the main application for monitoring sports training from the perspective of technical monitoring is the sports video analysis system [4–7]. However, the motion video analysis system still has serious shortcomings, mainly due to the inability of timely feedback and complicated operations. The test data can be obtained only one or two months or probably three or four months after the image is taken. Hence, the real-time monitoring cannot be realized as the obtained data has no practical guiding value. The bottleneck of rapid feedback of kinematics research lies in the method of data collection and processing. The traditional analytical method is to manually interpret the joints of the human body. The workload is huge, which seriously affects the speed of feedback. Moreover, the manual recognition operation is boring and often affected by errors experienced during fatigue and various operations along with poor repeatability. The operation of motion video analysis system is also relatively complicated, and it requires certain basic knowledge and strict training to operate it. For busy coaches, the operability is very poor.

In order to solve the problem of slow information feedback and poor operability of the sports video analysis system, this article takes sprint as the research problem, uses computer as the tool, and adopts the literature and expert interviews into account for measurement, software engineering modeling of sports [8–10], and human body construction. Research methods such as simulation and mathematical modeling transform the video into a jitter-free

digital image and automatically recognize the joint points of the human body. Typically, this is achieved by extracting the contour line of the moving human body, dividing the movement phase, determining the length of the scale, tracking the area, determining the joint point, performing special judgment, and processing of the situation. At the end, a computer application program "automatic recognition software system for human body joint points in sprint [11, 12]" is generated. The computer automatically recognizes the human body joint points in the sprint motion image, so that the motion video analysis system can accurately and timely feed back the motion information with good operability.

The method of using motion recognition technology to analyze the technical actions of sprinters during the exercise to improve the quality of training has gradually attracted people's attention. Through the recognition of exercise videos during the training process, people's motion status can be grasped, and computer analysis and recognition are used at the same time. The related parameters which are obtained reflect more intuitively the degree of standardization of the athletes' training movements and help coaches and athletes to analyze technical movements and find problems, thereby improving the quality of athletes' training.

The system designed in this article adopts a client/server design architecture. The client data acquisition module acquires and displays real-time video and transmits it to the server synchronously. After the server receives the data, it performs action recognition and feeds back the recognition result. The client and server use multithreading technology for data synchronization. The main contributions of this paper are as follows:

(1) Aiming at the shortcomings of traditional dual-stream convolutional neural networks in processing long-term video information, the time-segmented dual-stream network based on sparse sampling used in this paper can better express the characteristics of long-term motion. First, the continuous video frame data is divided into multiple segments, and a short sequence of data containing user actions is formed by randomly sampling each segment of the video frame sequence, which is then applied to the dual-stream network for feature extraction.

(2) This paper uses the Lucas–Kanade algorithm to extract the optical flow image because it contains the movement and time information of the target, which can effectively represent the movement of the pixels in different areas of the continuous frame image.

(3) This paper proposes a variety of data enhancement methods and network pretraining strategies to alleviate the risk of overfitting in the network training process.

(4) Analyzing the feature fusion methods of multiple dual-stream networks, network fusion at the convolutional layer is adopted and a three-dimensional convolution and pooling operation is used to perform feature aggregation operations, so that the network can express the spatiotemporal characteristics of actions more efficiently. Hence, it ensures higher recognition accuracy.

The rest of the paper is organized as follows. In Section 2, the background information is provided. In Section 3, the research methodology of our proposed work is explained. In this section, first, we discuss the system design followed by video action recognition. In Section 4, we provide comprehensive experimental results. Finally, we conclude the paper with future research directions in Section 5.

2. Background

In recent years, in the field of motion recognition, commercial institutions and scientific research institutes have achieved extremely important results. According to the data acquisition method for classification, the commonly used action recognition methods can be divided into two types based on wearable devices and vision. In sports such as table tennis, since it is a ball game dominated by arm movements, long-term exercise monitoring and analysis using wearable devices will not be conducive to the training effect of athletes. Therefore, the motion recognition method based on machine vision [9] is more suitable for this application' system development.

One of the important sources of human perception about the external environment is visual information. Human action recognition involves many fields such as machine vision, feature selection, and pattern recognition. It is a challenging advanced processing method for motion visual analysis [10]. In recent years, video-based action recognition technology has important research significance both in scientific research and in practical applications. Despite the efforts of researchers worldwide, though the use of machine learning to model and analyze human behavior has made great progress, they are also aware at the same time that the development of behavior recognition technology is still arduous and there are many problems that need to be solved urgently, which mainly include the following:

(1) The influence of the background factors of video environment: It is often difficult to avoid having a more complicated background in a video. The cluttered background factors will have varying degrees of influence on the subject of the action. Therefore, it is more difficult to identify and analyze the action in this case. In addition, factors such as changes in lighting in the video, whether there is occlusion, and the deviation of the viewing angle will also affect the feature performance of behaviors to varying degrees. The feature extraction and analysis of behaviors will become more complicated, resulting in the final recognition result getting difficult to achieve the desired accuracy.

(2) Differences in data within and between classes: Different subjects are doing the same behavior, and there are still certain differences in posture and scale. Therefore, for more complex actions, the data difference within the class makes the action recognition more difficult. In order to effectively improve the effect of behavioral action recognition, it is necessary to fully consider the impact of data changes within and between classes.

(3) Difference between the video action database and the real video data: At present, in order to facilitate effective training and effect verification of various research methods, most of the research is carried out based on some public video action databases. However, compared with various types of video data collected in real life, the video size, the quality, and other aspects are different, and there is a lack of rich changes.

In recent years, research community has poured a lot of enthusiasm for exploration around the abovementioned issues. With the rapid development of machine vision technology, the recognition and analysis of human actions has moved from a pioneering method of artificial representation to a research field based on deep learning methods. It has been able to learn from millions of videos and is applicable to almost advanced solutions for all daily activities. Looking at the research and development of video behavior and action recognition algorithms [11] in recent years, it can be divided into two categories according to the nature of the extracted behavior and action features, i.e., learning methods based on shallow features and methods based on deep learning.

In addition, because the convolutional neural network [13–16] only has a good expression of the two-dimensional features of the image, it is difficult to meet the extraction of the timing information in the video action. It can effectively combine the features of the previous moment to ensure that the network is able to understand the timing information. Since the video behavior can be regarded as a set of continuous and related image sequences, a dual-stream network model framework of CNN [17–19] combined with LSTM is produced. The main objective is to use the convolutional neural network to extract the spatial characteristics of the video frame image, to capture the timing information between actions through the RNN network, and to combine the temporal and spatial information [12] to recognize the video actions. In addition, the movement of the human body can be described by the movement of the skeleton joint points. Hence, a dual-stream RNN is used for spatial position and time dynamic characteristics of the human skeleton joint points for modeling. The RNN-based method can directly pass video frames as input to the network to realize an end-to-end action recognition.

3. Methodology

In this section, first, we discuss our system design in Section 3.1 followed by video action recognition of the athlete (sprinter in this case) in Section 3.2. Finally, we discuss the network model of our action recognition for the sprinter in Section 3.3.

3.1. System Design. Our system design is based on the client–server (C/S) architecture. The client and server are connected through a local area network, and the socket is used to realize the data communication between the front and back ends. The system assigns the data calculation and processing to the algorithm server to complete, and the client is mainly used as the carrier of display and control. This design method reduces the overhead of the client's storage and computing capabilities and effectively improves the real-time performance of the system. In addition, the design concept of multiclient sharing algorithm server is helpful for the expansion of system cluster design. The overall structure of the system is shown in Figure 1.

In this figure, the client program is mainly responsible for the display and collection of action video data. The real-time video stream is captured by the video data acquisition module (as shown in Figure 2). The data transceiver module transmits real-time video data to the server and receives the results of the data recognition processing on the server side. The display module is responsible for displaying the athletes' real-time training actions, videos, and corresponding recognition results. The server-side program is mainly responsible for recognizing the video data of the Ping-Pong ball receiving and serving action and synchronizing the recognition results. The data transceiver module on the server side receives the collected video data and synchronizes the results to the client. The action recognition module is responsible for feature extraction and action recognition of the video data. At the same time, the display module on the server side synchronizes the working status of the server in real time.

The detailed workflow of the system is shown in Figure 3. A thread is opened through the system client to monitor user control instructions and parse them. When receiving the instruction to start training, call the web camera to capture a real-time video stream containing the act of receiving and serving the ball. The client's data transceiver module sends the synchronization instructions and data to the server, notifies it to receive the video data, and calls the identification module for processing and identification. At the same time, the client program will also monitor and display the training video captured by the camera in real time. During user training, the client uses this thread to receive user instructions. If the thread receives an instruction to stop training before the end of the training, the client will stop the acquisition of the video stream and synchronize the message to the server. After the data transceiver module of the system server receives the video data transmitted by the client, it calls the algorithm program of the action recognition module to process the data and finally feeds back the recognition result to the client. At the same time, the server-side display interface will monitor the running status of the server in real time. The client data transceiver module will always monitor the server port and wait to receive the recognition result feedback by the server, call the training pattern in the database for comparison with the recognition

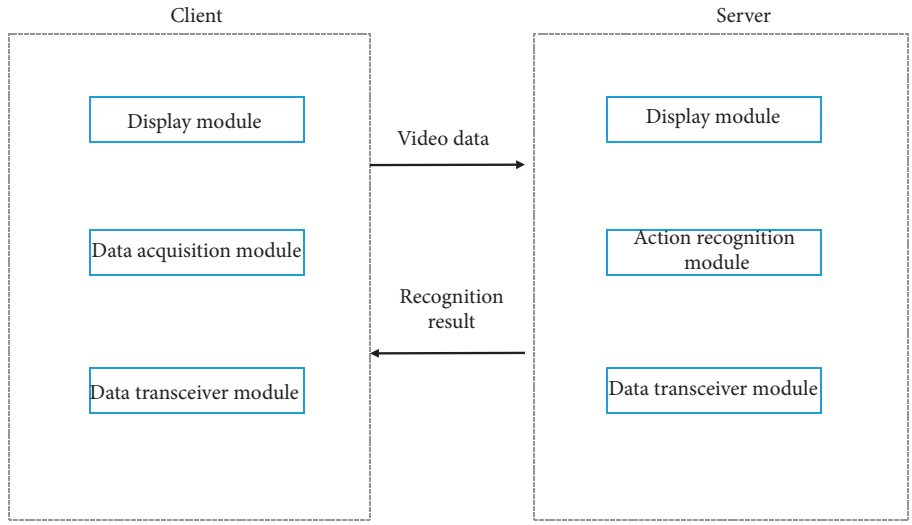

Figure 1: The overall structure of the system.

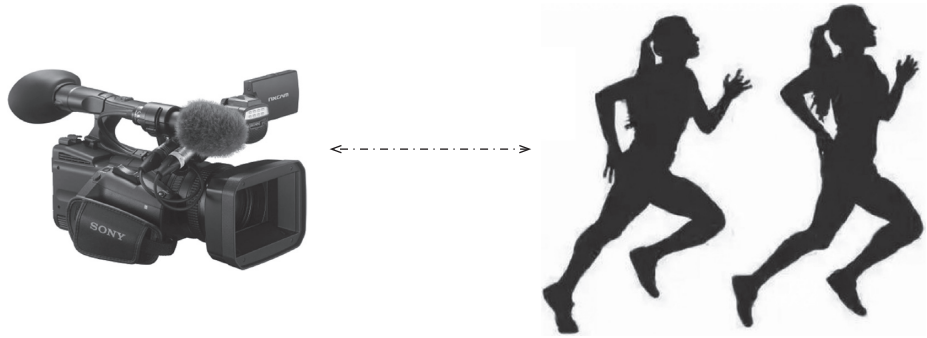

Figure 2: Real-time sprint video data collection.

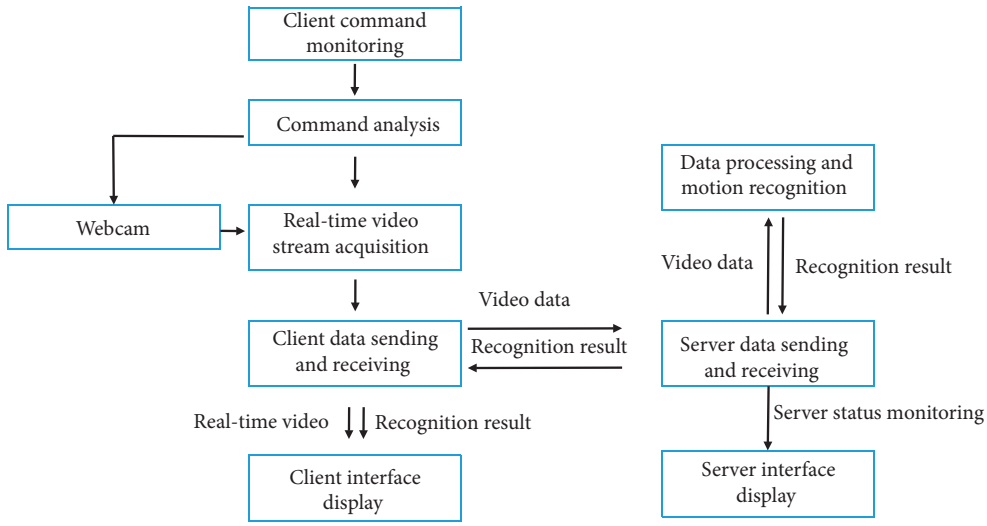

Figure 3: Detailed workflow of the system.

result, and display the completion of the training plan in real time on the interface, and the result will be displayed. It is stored in the database and used as a reference basis for coaches to guide and make training plans in the future.

3.2. Video Action Recognition. In this section, first, we discuss the dual-stream convolutional neural network (CNN) followed by optical flow feature extraction of sprinter's action video data.

3.2.1. Dual-Stream Convolutional Neural Network. In recent years, most of the research work has been inspired by the dual-stream convolutional neural network, which combines the spatiotemporal information extracted from the RGB image and the optical flow image of the video and extracts two types of features through two separate convolutional neural networks. We need to identify and produce the final prediction result. Video information contains two parts: space information and time information. Compared with static images, the time sequence component of video provides additional motion information that represents time for action recognition. The spatial information in the video is the position on each frame of the image, which represents the spatial information such as the target and the scene in the video; the time information refers to the change between video frames and carries the target movement information between consecutive video frames, including the movement of the camera or movement information of the target object, etc. The idea of realizing video action recognition mainly includes two categories, namely, the method of extracting video spatiotemporal features for video recognition and the method of retraining by using human skeleton node information as network input data. In this paper, the system uses ordinary camera equipment to collect motion video for recognition, so the dual-stream convolutional neural network is mainly used to extract video spatiotemporal features for action recognition and analysis.

As shown in Figure 4, the dual-stream convolutional neural network is composed of a convolutional network that expresses two-dimensional information of spatial flow and temporal flow and is used to process the spatiotemporal information of video data. This dual-stream network design architecture comes from biological vision. The spatial-stream network takes a single-frame RGB image of the video as input. The decoupling of the spatial stream and the temporal stream network also enables the use of image data on large image data sets. The pretraining of the spatial-stream network is used to recognize the surface features related to the action and realize the feature description in the spatial domain of the video. The spatial flow network is the same as the common static image recognition network, while the time flow network is to input multiframe stacked optical flow images into the network for training. It is used to learn the time features contained in the action, such as the movement and deformation of the target. For feature description of the video action in the time domain, we use the method of multitask training, which provides two softmax output layers for fusion. The output of the softmax layer is the probability of identifying the action category. Providing two softmax outputs is equivalent to the process of regularization. There are two main fusion methods: averaging and retraining an SVM classifier using the softmax layer for recognition. In short, the processing method of decoupling spatiotemporal features of dual-stream convolutional neural network better describes the movement information of sprinters.

The dual-stream convolutional neural network actually draws on the dual-stream pathway of information processing in the human brain's visual system, which can effectively capture the local motion information of the target in the video and improve the single-stream convolutional neural network's ability to solve the action recognition problem. However, there are also some shortcomings. First, because the prediction of the action video is obtained by the average prediction of the sampled video clips, the medium- and long-term time information is still lacking in the learned features. Secondly, since the training data samples are obtained by randomly sampling video clips, there may be a problem of incorrect assignment of labels for each category of data. Finally, the use of dual-stream convolutional neural networks requires the optical flow image to be precalculated and saved as the input of one of the networks. The network training requires separate training for single-frame RGB images and multiframe stacked optical flow images, so the network is difficult to achieve an end-to-end training.

3.2.2. Optical Flow Feature Extraction of Sprinter's Action Video Data. Optical flow represents the speed vector of the object in the video, including the instantaneous motion direction and speed of the pixel information. It can be used to represent the motion information of the sprinter in the video. The optical flow feature combines the static information and motion information of the image and is a good feature performance for describing the behavior of sprinters. This section will describe the optical flow feature extraction algorithm in detail by introducing the design ideas of the client real-time video data acquisition module. After the data is synchronized to the server through network communication, the optical flow image is extracted from it for sprinter action feature extraction and data processing process. Data preprocessing performed by data enhancement technology can effectively avoid the risk of overfitting during network training due to the small action dataset.

The essence of object movement is the relative positional movement between the object and the scene. When the observer observes the moving object, the scene of the moving object will form a continuously changing image on the retina of the eye, which seems to flow through the retina of the eye; hence, it is called optical flow. Optical flow expresses the instantaneous velocity field of image pixels, including the position, instantaneous velocity, direction, and other pieces of information of a certain pixel of the moving target in the video. The optical flow method is used to extract the motion information of the foreground target in the video, and the motion foreground can be monitored independently without knowing any information of the scene in advance. Optical flow is calculated mainly to determine the change of the same pixel in two adjacent frames by estimating the time domain changes of pixels in different areas of the two frames of images and the correlation between pixels. Next, the vector of the moving target in a single image field is obtained to realize the motion estimation of objects in the image.

The two most famous methods for calculating optical flow are the Hom–Schunck algorithm and the Lucas–Kanade algorithm. The Hom–Schunck algorithm is based on the fact that the gray level of the object image

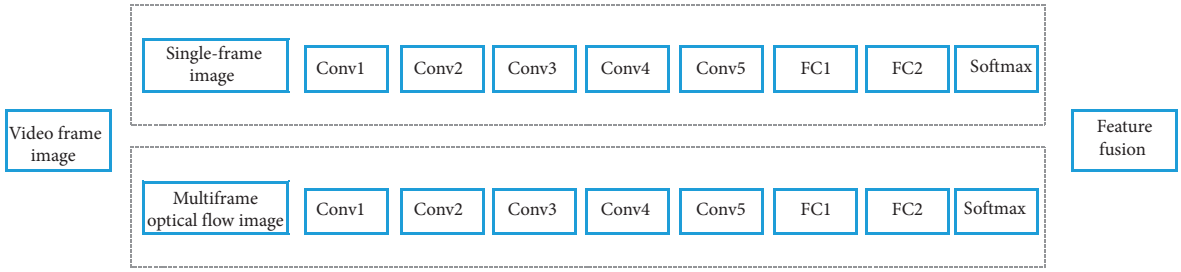

Figure 4: Dual-stream convolutional neural network model.

remains unchanged within a short time interval, assuming that the velocity vector field changes slowly in a given neighborhood, and the global smoothness constraint of the optical flow field is proposed. However, due to the smoothness assumption that the algorithm is based on, there may be large errors in the vector estimation of the optical flow in the edge area of the image or in the presence of occluded areas. The Lucas–Kanade algorithm uses local smoothness constraints, i.e., by assuming that all pixels in a small neighborhood have similar motions, and then realizes the estimation of optical flow. Compared with the previous algorithm, the Lucas–Kanade algorithm is simpler to implement and has lower computational complexity. Therefore, the system uses the Lucas–Kanade algorithm for the calculation and estimation of optical flow. The realization of this algorithm first needs to satisfy several assumptions:

(1) The outside has a constant intensity of light to avoid the intensity of the outside light from causing changes in the pixel value of the same point in the image.

(2) The time between adjacent frames of images is short enough so that the pixel difference between frames can be ignored when considering motion changes.

(3) The image pixels of the moving target perform similar movements in the same neighborhood.

Assuming that the brightness of the pixel at the point (x, y) at time$t$ is $I(x, y, t)$, the pixel point moves to the position $(x + \Delta x, y + \Delta y)$ at time $t + \Delta t$. The brightness at time t is $I(x + \Delta x, y + \Delta y, t + \Delta t)$. Based on the above assumption, the brightness of the pixel at the same point is constant; hence,

$$I(x, y, t) = I(x + \mathrm{d}x, y + \mathrm{d}y, t + \mathrm{d}t). \tag{1}$$

According to the assumption that the movement time of the target object is very short, the above equation can be Taylor-expanded to obtain the following equation:

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + \Delta = 0. \tag{2}$$

Ignoring the higher-order term Δ and deriving the above formula to Δt, the constraint formula for optical flow can be obtained as follows:

$$I_x\mu + I_y v + I_t = 0, \tag{3}$$

where I_x, I_y, and I_t can be calculated directly from the image, and (μ, v) is the optical flow of $I(x, y, t)$, which is an unknown quantity, and the optical flow value cannot be obtained only by the optical flow equation. Therefore, the selected Lucas–Kanade algorithm will solve the equation by attaching various optical flow constraints. This method is based on local smoothness constraints; that is, it is assumed that the pixels of the image have small displacements between adjacent frames and maintain approximate motion in the neighborhood. Then, the basic optical flow equation of the pixels in the neighborhood can be solved by the least square method.

According to the assumption of the Lucas–Kanade algorithm, all pixels in a small local area have similar motions. Based on the basic optical flow equation, an overdetermined equation can be obtained as follows:

$$I_x(X_i)\mu + I_y(X_i)v + I_t(X_i) = 0. \tag{4}$$

The optical flow estimation error is defined as

$$E = \sum_{X_i \in \theta} \left(I_x(X_i)\mu + I_y(X_i)v + I_t(X_i)\right)^2. \tag{5}$$

The optical flow calculation equation can be obtained by fitting by the least square method:

$$A^T A\vec{v} = A^T(-b), \tag{6}$$

where A is the coefficient matrix and $b = -(I_t(X_1), \ldots, I_t(X_n))^T$ is the constant term. When $A^T A$ is a nonsingular matrix, the solution of the above equation is

$$\begin{bmatrix} \mu \\ v \end{bmatrix} = \begin{bmatrix} \sum_i I_x(X_i)^2 & \sum_i I_x(X_i)I_y(X_i) \\ \sum_i I_x(X_i)I_y(X_i) & \sum_i I_x(X_i)^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i I_x(X_i)I_t(X_i) \\ -\sum_i I_x(X_i)I_t(X_i) \end{bmatrix}. \tag{7}$$

From this, the value of μ,v can be obtained.

3.3. Action Recognition Network Model. This paper uses a time-segmented convolutional neural network based on a sparse sampling strategy. Its structure is shown in Figure 5. The time-segmentation convolutional neural network used is to segment the entire video and sparsely sample short segments as the network input and extract the temporal characteristics of the optical flow image and the spatial characteristics of the RGB image to perform action recognition tasks. The time segmentation convolutional neural network first divides a video containing an action into several equal parts and then randomly extracts a short sequence from it that can effectively express the motion information in the entire video. For each sampled segment, feature extraction is performed through a dual-stream convolutional neural network. The temporal stream network captures the temporal structure information of the video and the spatial appearance information of the image captured by the spatial-stream network. Next, a corresponding dual-stream network prediction is generated for each short segment. Finally, an aggregation function is used to fuse the characteristics of time flow and spatial flow network as the recognition result of the entire video. This method can effectively extract the long-term information of the entire video, which is more accurate and effective than the recognition method of densely sampling the entire video segment; also, it no longer increases the computation cost. In the learning process, the loss value of the entire video prediction is optimized through the iterative update parameter calculation to realize the end-to-end network training process.

For a given action video frame data, first, divide it into N video frame sequences of equal length $\{S_1, S_2, \ldots, S_N\}$, and then randomly sample a short-term video sequence from each part S_N. The network models the short-term sequences $(T_1, T_2, \ldots, T_N)$ extracted from each part as the input number, corresponding to the equation below:

$$T(T_1, T_2, \ldots, T_N) = H(G(F(T_1, W), F(T_2, W), \ldots, F(T_N, W))). \tag{8}$$

4. Experimental Setup and Results

In this section, we discuss the experimental environment followed by the dataset that is used in this research. Moreover, we also talk about the training process of the model and the preliminary simulation we have obtained so far.

4.1. Experimental Environment. Since the experiment in this article needs to train a deep neural network, hence the scale is large, the structure is more complex, and the calculation scale is massive. The neural network training process needs to use GPU to accelerate the calculation. The experimental environment configuration is shown in Table 1.

In our experimental work, the data set is a 100-meter short film project, with a total of 5000 videos, each video action lasts about 2–10 seconds, the resolution is $320 * 240$, and the frame rate is 30 frames per second.

4.2. Training. The learning of network parameters is achieved by a small batch of gradient descent algorithm, with the batch size set to 256 and the momentum set to 0.9. Gradient descent in small batches can be accelerated by calculation of matrices and vectors, and the variance of updated parameters can be reduced to obtain more stable convergence. Using a batch, each time can reduce the number of iterations of convergence and at the same time make the result of convergence closer to the effect of gradient descent. For traditional gradient descent algorithms, if the actual objective function plane is a partially concave surface, then a negative gradient will make it point to a steeper position. This situation near the local optimal value of the objective function will cause the convergence rate to slow. At this time, it is necessary to give the gradient a momentum, so that it can jump out of the local optimum and continue to optimize in the direction of gradient descent, so that the network model can more easily converge to the global optimum.

For the time segmentation dual-stream network used in this system, set the initial weight of the spatial stream convolutional neural network to 1 and the initial weight of the time stream convolutional network to 1.5. The learning rate of the network training is set smaller: set the initial value of the spatial flow convolutional neural network to 0.01 and adjust it to one-tenth every 2000 iterations; set the initial value of the time flow convolutional neural network to 0.005; adjust it to one-tenth after 12000 and 18000 iterations of network parameters. In addition, the total time-consuming data training is as follows: the spatial streaming network requires about 2 hours and the time streaming network requires about 11 hours.

4.3. Experimental Results. The value of spatial flow convolutional neural network is set to 0.01 and adjusted it to one-tenth after 2000 iterations. Next, we set the initial value of the time flow convolutional neural network to 0.005 and adjust it to one-tenth after 12000 and 18000 iterations of network parameters. In addition, the total time-consuming data training is as follows: the spatial streaming network requires about 2 hours and the time streaming network requires about 11 hours.

In Figure 6 and Table 2, the accuracy for action sprint is computed. It can be seen from these results that separate spatial stream and time stream convolutional neural networks are far less effective in recognition of actions than the dual-stream fusion network model, and the recognition accuracy is about 5% to 19% lower. For different training methods, due to the small data set of the action category, the dual-stream network trained from scratch has the problem of overfitting. Hence, the performance is the worst, and the pretraining of the spatial-stream network and the cross mode of the time-stream network is time consuming. The dual-stream convolutional neural network initialized by pretraining has a better recognition effect, and the recognition rate can reach 94.3%, which also means that this pretraining method can effectively reduce the risk of overfitting.

Figures 7 and 8 show the recognition results on different frames and the impact of different input resolutions on model performance.

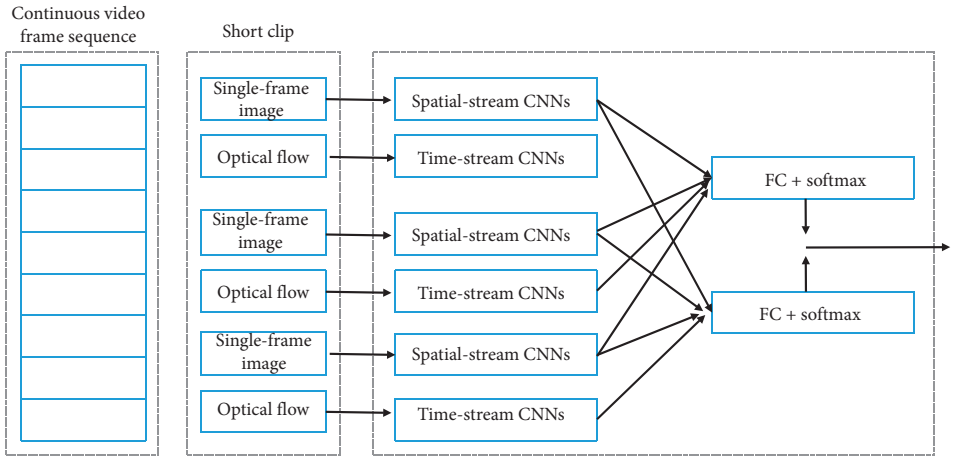

Figure 5: Action recognition network model.

Table 1: Experimental hardware platform and software simulation environment.

CPU	Intel® Core™ i5-4200M CPU @ 2.50 GHz
RAM	16.00 GB
Operating system	Centos 8.0
Development environment	PyCharm 2020.2
Programming language	Python 3.6.5

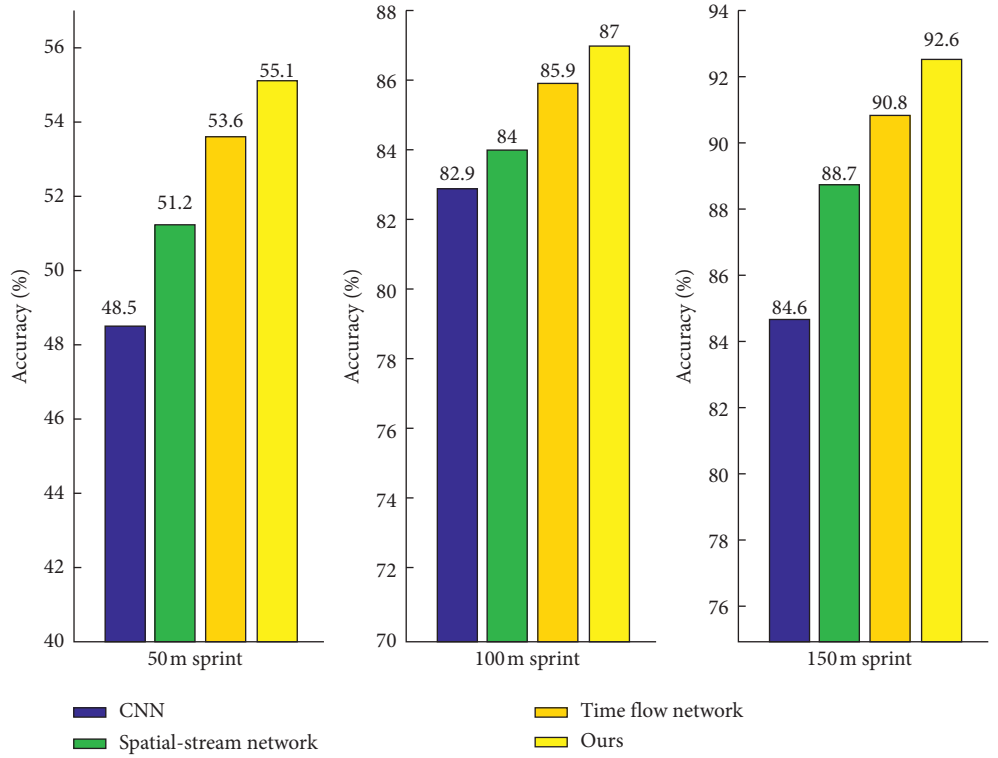

Figure 6: Action recognition network model.

Table 2: Experimental results.

Type	Spatial-stream network (%)	Time flow network (%)	Dual stream (%)
Classic	72.6	73.5	85.6
Spatial flow	75.2	81.6	88.7
Cross mode	77.1	82.6	92.6

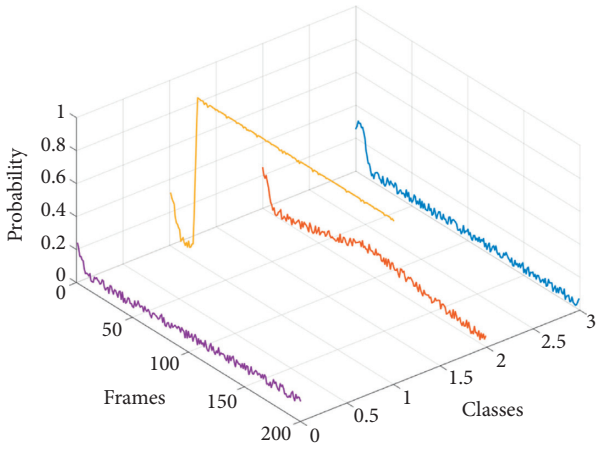

Figure 7: Recognition results on different frames.

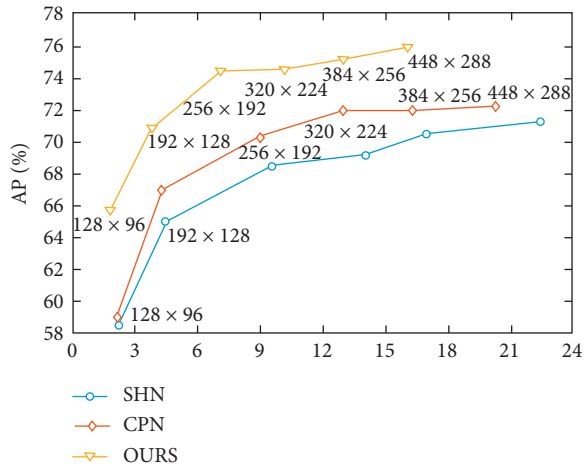

Figure 8: The impact of different input resolutions on model performance.

5. Conclusion

This paper takes sprint as the research object and constructs the image of sprinter's action recognition based on machine learning. Then, in view of the shortcomings of the traditional dual-stream convolutional neural network for processing long-term video information, the time-segmented dual-stream network based on sparse sampling is used that can better express the characteristics of long-term motion. First, the continuous video frame data is divided into multiple segments, and a short sequence of data containing user actions is formed by randomly sampling each segment of the video frame sequence, and then it is applied to the dual-stream network for feature extraction. The optical flow image extraction involved in the dual-stream network is implemented using the Lucas–Kanade algorithm. Optical flow contains the mobility information of the target, which can effectively represent the movement of pixels in different areas of the continuous frame of image. Due to the small amount of action video data (test data), this paper adopts numerous data enhancement methods and network pre-training strategies during data processing to alleviate the risk of overfitting in the network training process. The system in this paper has been tested in actual scenarios, and the results show that the system design meets the expected requirements.

Data Availability

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

Conflicts of Interest

All the authors declare no conflicts of interest.

Acknowledgments

This work was supported by Jiangxi University of Science and Technology Philosophy and Social Science Prosperity Plan Cultivation Project "Research on the Construction of Mountain Sports Town in Jiangxi Province" (Grant no. FZ18-YB-18).

References

[1] T. Haugen, S. Seiler, Ø. Sandbakk, and E. Tønnessen, "The training and development of elite sprint performance: an integration of scientific and best practice literature," *Sports Medicine-Open*, vol. 5, no. 1, pp. 1–16, 2019.

[2] H. Liang, "Evaluation of fitness state of sports training based on self-organizing neural network," *Neural Computing and Applications*, vol. 33, pp. 1–13, 2021.

[3] C. M. Fairman, M. C. Zourdos, E. R. Helms, and B. C. Focht, "A scientific rationale to improve resistance training prescription in exercise oncology," *Sports Medicine*, vol. 47, no. 8, pp. 1457–1465, 2017.

[4] A. Lopatiev, O. Ivashchenko, O. Khudoliy, Y. Pjanylo, S. Chernenko, and T. Yermakova, "Systemic approach and mathematical modeling in physical education and sports," 2017.

[5] S. Subhash, T. Obafemi-Ajayi, D. Goodman, D. Wunsch II, and G. R. Olbricht, "Predictive modeling of sports-related concussions using clinical assessment metrics," in *Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 513–520, IEEE, Canberra, Australia, December 2020.

[6] A. Laredo, L. P. Concepcion, and N. Mamani-Macedo, "A proposal for modeling of the management of talent recruitment and training in Peruvian sports centers," in *Proceedings*

of the 5th Brazilian Technology Symposium, pp. 505–513, Lima, Peru, November 2021.

[7] A. L. Clouthier, G. B. Ross, and R. B. Graham, "Sensor data required for automatic recognition of athletic tasks using deep neural networks," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 473, 2020.

[8] K. Choroś, "Highlights extraction in sports videos based on automatic posture and gesture recognition," in *Proceedings of the Asian Conference on Intelligent Information and Database Systems*, pp. 619–628, Kanazawa, Japan, April 2017.

[9] G. Thomas, R. Gade, T. B. Moeslund, P. Carr, and A. Hilton, "Computer vision for sports: current applications and research topics," *Computer Vision and Image Understanding*, vol. 159, pp. 3–18, 2017.

[10] A. Filgueiras, E. F. Quintas Conde, and C. R. Hall, "The neural basis of kinesthetic and visual imagery in sports: an ALE meta–analysis," *Brain Imaging and Behavior*, vol. 12, no. 5, pp. 1513–1523, 2018.

[11] Y. Kong and Y. Fu, "Human action recognition and prediction: a survey," 2018, http://arxiv.org/abs/06.11230.

[12] P. E. Martin, J. Benois-Pineau, R. Péteri, and J. Morlier, "Sport action recognition with siamese spatio-temporal cnns: application to table tennis," in *Proceedings of the 2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, IEEE, La Rochelle, France, September 2018.

[13] H. C. Shih, "A survey of content-aware video analysis for sports," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1212–1231, 2017.

[14] L. I. U. Ziyu, "Application of college basketball training teaching based on sports video analysis under Network Multimedia," *Solid State Technology*, vol. 64, no. 1, pp. 170–181, 2021.

[15] C. Ning, "Design and research of motion video image analysis system in sports training," *Multimedia Tools and Applications*, pp. 1–19, 2019, In Press.

[16] Y. Li, "Research on sports video image analysis based on the fuzzy clustering algorithm," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 6630130, , 2021.

[17] W. Cai, B. Liu, Z. Wei et al., "TARDB-Net: triple-attention guided residual dense and BiLSTM networks for hyperspectral image classification," *Multimedia Tools and Applications*, vol. 80, pp. 11291–11312, 2021.

[18] X. Zhang, Y. Yang, Z. Li, X. Ning, Y. Qin, and W. Cai, "An improved encoder-decoder network based on strip pool method applied to segmentation of farmland vacancy field," *Entropy*, vol. 23, no. 4, p. 435, 2021.

[19] X. Ning, X. Wang, S. Xu et al., "A review of research on co-training," *Concurrency and Computation: Practice and Experience*, Article ID e6276, 2021.