IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON IEEE SYSTEMS, MAN AND CYBERNETICS SOCIETY SECTION

Received 15 October 2022, accepted 3 November 2022, date of publication 11 November 2022,
date of current version 6 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3221530

RESEARCH ARTICLE

The Application of Deep Convolution Neural Network in Volleyball Video Behavior Recognition

CHEN LIANG AND ZHIJUN LIANG
China Volleyball College, Beijing Sport University, Beijing 100084, China
Corresponding author: Zhijun Liang (liangzhijun@bsu.edu.cn)

ABSTRACT The purpose is to minimize subjective errors in the manual analysis of volleyball game videos and improve the traditional Human Behavior Recognition (HBR) algorithms' excessive calculation, high hardware requirements, and poor long-stream video modeling ability. Firstly, this paper expounds on the relevant theories. Secondly, a fusion Convolutional Neural Network (CNN)-based HBR model is implemented. It combines the two-stream CNN (TSCNN), Three-Dimensional (3D) CNN, and Long Short-Term Memory (LSTM) and gives full play to the LSTM's long-term Dynamic Information Extraction (DIE) ability. Finally, the public dataset is selected to verify the model's volleyball-game-video-oriented HBR performance. Here are the experimental results. (1) The optimal key parameters of the proposed fusion-CNN-based HBR model are determined as follows: the number of video segments is three, the average method is used for feature fusion, and then the HBR accuracy is the highest when the fusion ratio of spatial feature map and temporal feature map is 4:6, and the learning rate is 0.0014. (2) The average recognition accuracy of the proposed fusion-CNN-based HBR model on three different datasets is 4%, 2.7%, and 3% higher than other popular networks, respectively. The improvement effect of the model is remarkable, and it is suitable for studying Human Behavior Analysis (HBA) in volleyball match videos. Finally, the proposed HBR model can provide more accurate results for volleyball videos' HBR, which is significant in promoting the rapid development of volleyball sports. The proposed model can classify and label videos and understand and describe video behaviors to simplify video processing procedures and save social resources.

INDEX TERMS Convolutional neural network, behavior recognition, fusion network, volleyball video, recognition accuracy.

I. INTRODUCTION

In the last several decades, team ball games like volleyball have been among the top popular sports worldwide with successful commercialization [1]. Volleyball demands less physical contact than basketball, football, and baseball. For that reason, many people choose volleyball over basketball and football for physical exercise or entertainment [2]. Nevertheless, players must hit the ball with specific body parts and proper poses in volleyball to avoid injuries and improve serving and attacking accuracy and strength. Besides, the individual influence on game results might be more significant than other sports, given its relatively short innings, fewer players, and smaller courts [3]. During a match, the referee judges the athlete's behavior through direct observation or video replay, which is somewhat subjective. Thus, referees' experience and expertise might weigh too much sometimes. Therefore, the judgment might not always be accurate or comprehensive, and instability might become another concern. Volleyball game videos can be used to analyze athletes' individual and group behaviors. It is also very important for the players and coaches to learn from their mistakes and opponents' competencies. Compared with ordinary Two-Dimensional (2D) images, video information is

The associate editor coordinating the review of this manuscript and approving it for publication was Laura Celentano.

more straightforward and better reflects real-game scenarios. Therefore, the traditional 2D image-oriented feature extraction is unsuitable for video behavior recognition [4]. With the rapid development of Computer Technology (CT), especially Artificial Intelligence (AI), spatiotemporal feature extraction methods are maturing and prevailing.

Deep Learning technology is widely used in image processing. For example, Chen et al. studied applying an optimal evacuation model based on Deep Learning in public structure design [5]. Hu et al. explored the application of Deep Learning in medical image processing [6]. Liu et al. improved the accuracy of human abnormal behavior recognition through a two-stream Convolution Neural Network (CNN) and obtained a higher accuracy than other literature methods [7]. Hu proposed an improved spatiotemporal CNN to simplify the high-complexity population anomaly detection model, extract time-related features, and increase training samples. They discovered that most abnormal behaviors could be detected, and the alarm information could be transmitted in real-time [8]. Chen et al. constructed a behavior analysis system based on Three-Dimensional (3D)-CNN. A spatial-stream CNN was used to extracted spatial features, and a temporal-stream CNN captured human motion information. Compared with traditional 2D-CNN, 3D-CNN could comprehensively extract video behavior features from space and time dimensions [9]. Chen et al. built a lightweight and efficient 2D and 3D integrated model. In the integrated model, 2D-CNN could obtain the feature mapping of the input image, and 3D-CNN could deal with the time relationship between frames. The proposed integrated model used the advantages of 3D-CNN in video time modeling to reduce the model complexity. As a result, the proposed method presented excellent and faster performance [10]. Lan et al. put forward a Two-Stream CNN (TSCNN) model framework and analyzed the optimal parameters for Human Behavior Recognition (HBR). The proposed model extracted the appearance features of human behaviors from the spatial domain and the motion features of human behaviors from the temporal domain. Finally, the two streams' Softmax output was combined with linear weighting to realize HBR [11]. The attention-based model transformer performs better than Long Short-Term Memory (LSTM). Du et al. emphasized the Spatio-temporal transformation of video understanding [12].

TSCNN and 3DCNN are among the most widespread HBR methods based on the related literature review. However, the traditional TSCNN has poor modeling ability for long-stream videos. Meanwhile, the 3DCNN must be pre-trained with vast numbers of video samples, which is time-consuming and requires high-performance computer hardware. Besides, with the changing video images, other long-term dynamic information may play a key role in HBR. Against such concerns, this paper creatively combines the TSCNN, 3DCNN, and Long Short-Term Memory (LSTM) network with excellent long-term Dynamic Information Extraction (DIE) ability to build a fusion-CNN-based HBR model. The literature review section analyzes the research status of group behavior recognition. Firstly, the basic theory of CNN and LSTM is introduced. Secondly, the construction process of the fusion-CNN-based HBR model is described in detail. Finally, the video recognition behavior performance of the HBR model is verified by selecting public data sets. The model performance is evaluated. The HBR model based on the fusion-CNN proposed here can provide more accurate volleyball video behavior recognition. The finding is of great significance for promoting the rapid development of volleyball events. The practical significance is to achieve the goal of saving resources through intelligent recognition of human movements in volleyball videos. The main contribution is to promote the intelligent development of volleyball training.

II. RELATED RESEARCH

Group behavior recognition is more challenging than sensing individual behavior, and no optimal solution has been explored, being a relatively new research field. Currently, group behavior recognition algorithms can be divided into two categories. One uses manual features and learning frameworks, and the other uses Deep Learning technology. Traditional methods usually extract manual Spatio-temporal features, such as Motion Boundary Histograms (MBH) features and Histograms of Orientated Gradients (HOGs) features, and then use graphical models to identify group behaviors. As one of the earliest works in group behavior recognition, researchers defined crowd background by space-time descriptor, which used people's posture and speed to extract visual clues about a certain behavior performed by individuals in the crowd. Soon, some scholars used an adaptive potential learning algorithm structure to model the hierarchical relationship from human-level information to group-level interaction. In addition to time-space characteristics, tracking algorithms could also provide contextual information about personnel. Researchers proposed a unified model by combining the framework and tracking clues to conduct behavior recognition. Further, some scholars combined multi-target tracking with group behavior recognition and suggested using the bottom-up method, transmitting information about behavior from people's trajectory estimation. First, they obtained human posture from semantic attributes and then determined the interaction between people, namely the paired relationship between individuals. The behavior of atomic pairs described each interaction in this method. Finally, they illustrated group behavior by using interaction sets. Some scholars introduced a top-down method to represent all detected people in an undirected weighted graph. Each side of the diagram described the interaction between the two people.

With the development of Deep Learning, computer vision research has witnessed significant improvements, including image classification, Human Behavior Recognition (HBR), and video classification. For behavior recognition and video classification, early research only used features obtained

by the Convolutional Neural Network (CNN). Some recent studies combined CNNs and Recurrent Neural Networks (RNN), considering optical flow information and original RGB (Red, Green, Blue) information. Other methods mainly used TSCNN. Some studies still use regional information and TSCNN methods. However, these studies only use local information when identifying behaviors. So far, scholars have proposed various Deep Learning methods to identify group behavior. Some scholars proposed the combination of hierarchical graph models and used a multi-step information transfer method between neural network layers. Based on this approach, they developed a common framework by combining graphical models and deep networks with structural learning. Specifically, the RNN network was used for sequential reasoning. Some researchers also used the LSTM algorithm and head detection for motion recognition.

Although many excellent network structures have been proposed in group behavior recognition, these algorithms still have different problems. Indeed, various multi-agent interaction types have been considered. However, most interaction research, especially human-human interaction, is described by handmade features. The interaction descriptors often lack self-learning abilities and quantifiable methods, and the shallow features cannot encode high-level information. Handmade feature extraction approaches require strong prior knowledge and often lose the temporal relationship information. Therefore, their differentiation ability cannot be guaranteed, nor can they be improved in the datasets. The Deep Learning method has achieved good recognition performance in group behavior recognition. However, all the studies take group videos as a whole. Group behavior is not a simple superposition of individual behavior but comprehensive temporal-spatial interaction of individual behaviors. Thus, group behavior recognition overlooking individual interaction and behavior might not extract accurate individual features. The existing feature extraction algorithm cannot fully consider the individual's temporal and contextual information, failing to meet the structural requirements of individual behavior. Thus, the recognition accuracy is low. In the group recognition task, the hierarchical framework is widely used to represent the relationship between individuals and between individuals and corresponding groups and has achieved good recognition performance. However, the existing methods only apply the global features to the network framework, ignoring the relatively important features that enhance the global and local features. The attention mechanism is rarely used, even if it can help the algorithm focus on the essential information of the image or video. Based on the above analysis, relying on the LSTM's excellent long-term dynamic information extraction, this paper combines the TSCNN and 3DCNN to build an HBR model based on the fusion CNN.

III. METHOD OF MODEL CONSTRUCTION

This section will build an HBR model based on fusion CNN, namely the CNN + LSTM. Experiments are designed to verify the performance of the model. Firstly, the CNN and LSTM algorithms are analyzed and combined. Secondly, the behavior recognition model is implemented based on fused CNN. Finally, the model performance evaluation experiment is designed.

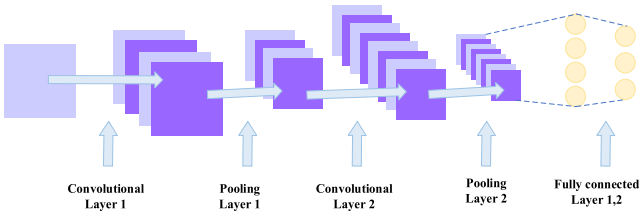

FIGURE 1. Typical CNN architecture.

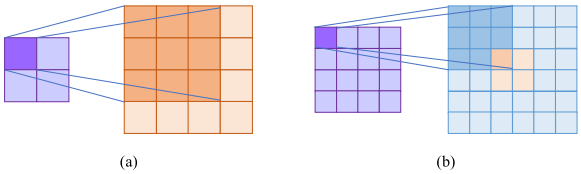

FIGURE 2. Operation of convolution and deconvolution (a. convolution; b. deconvolution).

A. RESEARCH THEORETICAL BASIS

CNN is a typical differentiated deep network structure based on minimizing the requirements of preprocessing data. It is one of the representative algorithms of Deep Learning and is used in the research fields of Computer Vision (CV) and Natural Language Processing (NLP) [13]. CNN is inspired by the human visual structure and is adaptable for supervised and unsupervised learning. Significantly, the parameter sharing of the convolution kernel in the hidden layer and the sparsity of inter-layer connection enable CNN to extract grid-like topology features with less computation [14]. Typically, a CNN comprises three parts: convolution, pooling, and fully connected. Among them, the convolution layer is mainly responsible for extracting the local and global features of the input data, and the pooling layer is accountable for reducing the magnitude of the parameters. Finally, the bottom parameters are mapped to the new space in the fully connected layer to collect the parameters and further calculate. Such a mechanism can quickly, effectively, and accurately obtain the data feature [15]. Fig. 1 sketches a common CNN architecture.

CNN's representational learning ability enables it to learn the internal characteristics of data and optimize the model structure and parameters. Generally, the shallow convolution layer at the front end of the CNN network can use a smaller perception domain to learn the local features, such as image texture. In comparison, the deeper convolutional layer at the back end uses a larger perception domain to learn abstract features, such as the size and orientation of objects in the image [16]. Eq. (1) calculates the convolution operation:

$$d(x, y) = \sigma\left(\sum_{N=0}^{N-1}\sum_{M=0}^{M-1} w^{N,M} u(x+n, y+m) + b\right) \tag{1}$$

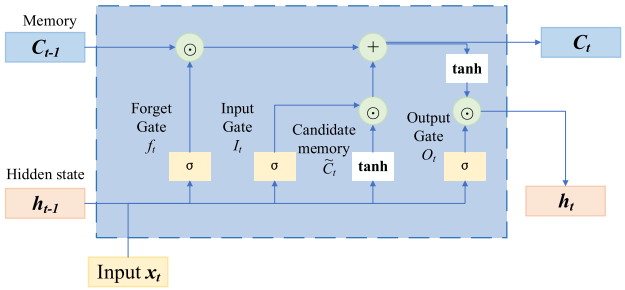

FIGURE 3. Example of LSTM structure.

In Eq. (1), $\sigma()$ indicates the Activation Function (AF). N and M represent the length and width of the convolution kernel, respectively. $w^{(N,M)}$ denotes the convolution kernel weight at the pixel (n, m). u and b stand for image feature and offset.

In CNN, extracted features can classify images or recognize and position objects. Nevertheless, CNN extracts the features around "a small neighborhood" using the perception domain. Thus, CNN cannot accurately fine-segment the image and recognize the specific object contour. Therefore, Full Convolutional Networks (FCN) came into being to solve the limitation of CNN. The most significant difference between FCN and CNN is to change the fully connected layer at the end of the network into a convolution layer. Images of any size can be input into FCN. Furthermore, an anti-convolution layer can be connected after the last convolution layer for upsampling to restore the image size output by the network to the same size as the input image [17]. Fig. 2 displays the convolution and deconvolution operations in upsampling.

LSTM is a Recurrent Neural Network (RNN) [18]. Nevertheless, the traditional RNN has a long-term dependence problem, which the LSTM network can solve through accumulation instead of continuous multiplication. Therefore, the derivative is also accumulative to avoid the gradient disappearance or explosion of the traditional RNN [19]. LSTM comprises four structures: cell state, forget gate, input gate, and output gate [20]. Fig. 3 gives an example of the LSTM structure.

1) The forget gate determines whether the information in the cell state is passed or filtered. Its input has two parts. The first is the output h_{t-1} of the last time, and the other is the input x_t of the current time. Forget gate outputs a vector f_t with each dimension between 0∼1, b_f is an offset. Eq. (2) is the specific calculation step:

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \tag{2}$$

2) The input gate determines whether the information at the current time will be added to the cell state. Its input, like the forget gate, is the output h_{t-1} of the previous time and the current time input x_t. Then, Eqs. (3) and (4) hold:

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C) \tag{4}$$

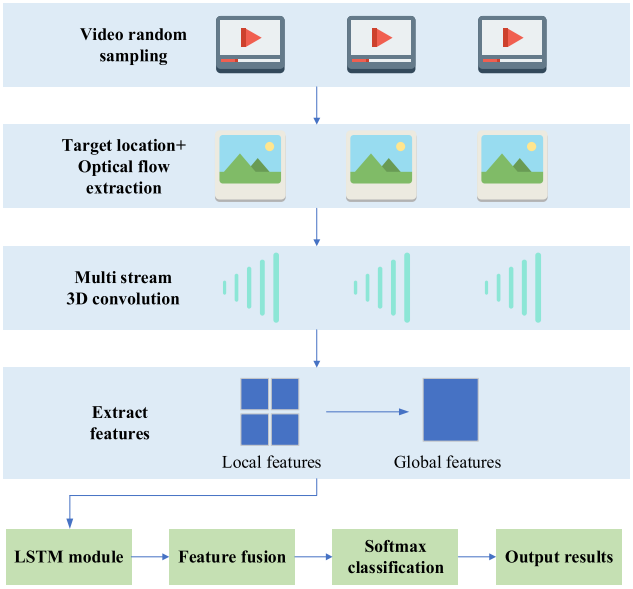

FIGURE 4. Structure of the proposed HBR model based on fusion CNN.

In Eqs. (3) and (4):
i_t: Control quantity of candidate cell state;
$\tilde{C}_t$: Candidate cell state;
C_t: Generated new cell state. Eq. (5) presents the specific calculation:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

3) The output gate determines the output at the current time. The output depends on the input of the previous time and the current time, and the new cell state o_t. The specific calculations are given in Eqs. (6) and (7):

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t * \tanh(C_t) \tag{7}$$

B. CONSTRUCTION OF BEHAVIOR RECOGNITION MODEL BASED ON FUSED CNN

Traditional CNN can only recognize each frame separately without considering the inter-frame motions of the time dimension, leading to low-quality video feature extraction [21]. Currently, TSCNN, 3DCNN, and LSTM are widely used in the research of video feature extraction. Both TSCNN and 3DCNN can extract spatial and temporal features in the video, but the traditional TSCNN has poor modeling ability for long-term video [22]. The 3DCNN needs to use a large amount of video data for pre-training, which is time-consuming and, thus, has high requirements for computer hardware [23]. Moreover, in time-variant video images, some other long-term dynamic information may be essential for HBR. Against the above problems, this paper proposes an HBR model based on fusion CNN. Firstly, a Multi-Stream 3DCNN (MS-3DCNN) module is proposed to extract local and global features of video images by combining

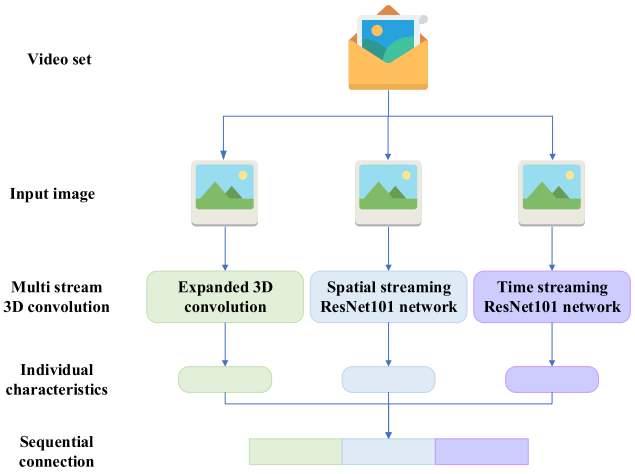

FIGURE 5. Structure of the MS-3DCNN module.

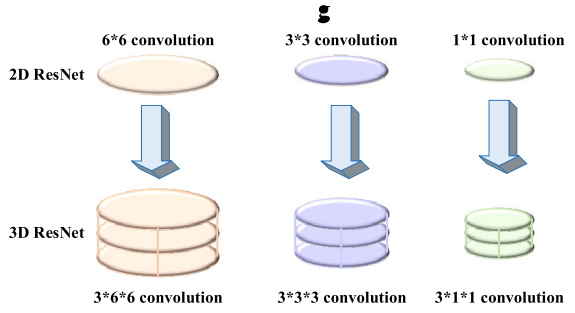

FIGURE 6. 2D expansion diagram.

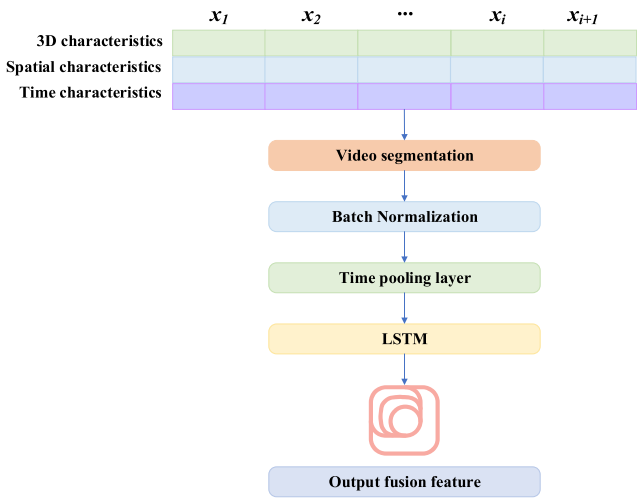

FIGURE 7. Structure of the LSTM feature segmentation module.

and improving TSCNN and 3DCNN. Then, the long-term dynamic information in the video is extracted by LSTM. The structure of the fusion-CNN-based HBR model is illustrated in Fig. 4.

In Fig.4, the model randomly samples the volleyball match video. At the same time, to reduce the calculation complexity and improve the recognition accuracy, an input sampling module, a target location extraction module, a Multi-Stream-3Dimensional Convolution Neural Network (MS-3DCNN) module, and a classification LSTM module are included. In the video behavior recognition task, the timing segmentation method is first used for the input video sequence [24]. The specific method is to divide the video sequence into several frames and segments, sample one image frame randomly in each segment, and then use the Faster R-CNN network to locate the people in the image. The image sequence comprises several frames before and after the person frame image, and the optical flow information feature map is obtained after the optical flow extraction. Then, the input image is divided into multi-frame image sequences. The sampled single-frame image and multi-frame optical flow diagram sequence are input into the MS-3DCNN, and the individual output features are connected to obtain the global features. After that, the output of the individual feature by each MS-3DCNN module is input into the segmented LSTM module. The output fusion features are fused with the global features again. Finally, the final individual behavior recognition result is obtained through the fully connected layer and Softmax classification operation. This paper presents a model based on 3D CNN and LSTM. In particular, 3D CNN is used to extract features from input image sequences to obtain feature sequences containing time information. LSTM processes the extracted feature sequences and uses the Softmax layer for classification. In the recognition model, after the CNN processes the data, the data will be adjusted to enter the LSTM. The feature vectors of each frame of the 3D CNN will fill the time series, and LSTM will perform recursive operations according to the time series. The result of each recursive operation is the fusion of all the previous and current features. In fact, LSTM obtains the time information between frames based on the spatial information of each frame extracted by the 3D convolutional network. Each module in the model will be briefly introduced below.

1) MULTI-STREAM-3DIMENSIONAL CONVOLUTION NEURAL NETWORK MODULE

The MS-3DCNN module mainly includes expanded 3DCNN, spatial stream Residual Neural Network (ResNet)101, and temporal stream ResNet101 network. Fig. 5 outlines the specific structure.

The deep neural network has some defects, such as easy falling into local minima, and the training results depend on initial random weights. The gradient explosion will cause network instability in a deep multilayer perceptron network. In the best case, the network cannot learn from the training data, and in the worst case, the weight value cannot be updated again. Gradient explosion leads to an unstable learning process. The neural network mainly maps the input data to high-dimensional space to better complete "data classification." Such nonlinear transformation can greatly improve data classification ability. The deeper the neural network is, the higher the accuracy is in a certain range. Still, with the continuous increase of network depth, the model training

convolutes, resulting in a gradual accuracy decline. Against such a dilemma, ResNet is proposed by adding a linear transformation branch to the traditional neural network and replacing the direct fitting basic mapping H(x) with F(x)+x to fit the residual mapping. The ''quick connection'' method eliminates the difficulty of deeper neural network training [25]. The residual connection mode is described by Eqs. (8) and (9):

$$y_i = x_i + F(x_i, W) \tag{8}$$

$$x_{i+1} = f(y_i) \tag{9}$$

In Eqs. (8) and (9), x_i and x_{i+1} are the input and output of layer i. $F(x_i, W)$ represents residual mapping. $f()$ indicates the Rectified Linear Unit (ReLU) function.

ResNet101 has shown excellent static image feature extraction ability in various research [26], so it is used as the benchmark network for spatial flow and temporal flow feature extraction. Then, a 3D CNN is obtained by expanding the traditional 2D ResNet101, as in Fig. 6.

The main expansion operations are as follows: firstly, the 2D parameters are copied directly into the 3D kernel along the time dimension. Then, three 2D convolution kernels are integrated as a 3D kernel to provide the multiple temporal dimension parameters. The above operations are described in Eqs. (10) and (11):

$$K_m^l = \left\{k_{m1}^l, k_{m2}^l, k_{m3}^l\right\} \tag{10}$$

$$K^l = C_l \sum_{i=0}^{m-1} K_i^l \tag{11}$$

In Eqs. (10) and (11), $k_{m1}^l, k_{m2}^l, k_{m3}^l$ represents the 2D convolution kernels of $m1, m2, m3$ channels in l layer. K_m^l indicates the corresponding 3D convolution kernel. C_l denotes the operation of fusing all 2Dconvolution kernels into one 3D convolution kernel K^l. The pre-trained 2D network mainly provides the parameters of the expanded 3D CNN during training. There is no need to use other datasets for pre-training, thus significantly saving time and computational cost.

2) LONG SHORT-TERM MEMORY FEATURE SEGMENTATION MODULE

LSTM network has a good performance in the long-term DIE for videos. Here, the LSTM network is connected with the time pooling layer in the MS-3D CNN module, and the LSTM feature segmentation module is constructed to extract the temporal dynamic information. The structure of the LSTM feature segmentation module is depicted in Fig. 7.

In the LSTM feature segmentation module, the input data are the serial feature sequence output by three networks in the MS-3DCNN module, divided into the same number of segments as the sampling stage. After batch normalization, the distinctive features of each segment are extracted from the serial feature sequence through the time pooling layer and, finally, input into the LSTM module for other long-term DIE.

3) FEATURE FUSION MODULE

The feature information extracted by the above modules needs to be fused in a certain order, mainly including spatial feature fusion and temporal feature fusion. First, the fusion function is set as $f : x_t^a, x_t^b \rightarrow y_t$. H, W, and D are the height, width, and the number of channels of the feature map, respectively. At the same time, $H_1 = H_2 = H_3$, $W_1 = W_2 = W_3$, and $D_1 = D_2 = D_3$ for different feature maps. Then, the two feature maps fused at time t can be expressed by Eq. (12):

$$x_t^a \in R^{H_1 \times W_1 \times D_1}, x_t^b \in R^{H_2 \times W_2 \times D_2} \rightarrow y_t \in R^{H_3 \times W_3 \times D_3} \tag{12}$$

Spatial feature fusion mainly includes four additive fusion methods, maximum value method, stacking method, and average value method, which will be described respectively below:

1) Additive fusion method

CNN randomly generates a channel number. The additive fusion method only needs to consider the relationship between corresponding networks. In this way, any communication relationship can be used to optimize the convolution kernel in the subsequent learning tasks. On the same spatial position and channels i, j, and j, the sum of the two feature maps is calculated by Eqs. (13) and (14):

$$y^{sum} = f^{sum}\left(x^a, x^b\right) \tag{13}$$

$$y_{i,j,d}^{sum} = x_{i,j,d}^a + x_{i,j,d}^b \tag{14}$$

In Eqs. (13) and (14): $1 \le i \le H$, $1 \le j \le W$, $1 \le d \le D$, $x^a, x^b, y \in R^{H_1 \times W_1 \times D_1}$.

2) Maximum value method

The maximum value method outputs the maximum value of two feature maps, where the corresponding relationship between channels is also randomly generated. The variables are described as Eqs. (15) and (16):

$$y^{max} = f^{max}\left(x^a, x^b\right) \tag{15}$$

$$y_{i,j,d}^{max} = max\left\{x_{i,j,d}^a, x_{i,j,d}^b\right\} \tag{16}$$

3) Stacking method

The stacking method is mainly to stack two feature maps at the same position of the same channel. Thus, the connection between the feature maps is in series, which is different from the additive fusion method. The corresponding variables are described in Eqs. (17) and (18):

$$y^{cat} = f^{cat}\left(x^a, x^b\right) \tag{17}$$

$$y_{i,j,2d}^{cat} = x_{i,j,d}^a, y_{i,j,2d-1}^{cat} = x_{i,j,d}^b \left(y \in R^{H \times W \times 2D}\right) \tag{18}$$

4) Average value method

The average value method averages the feature values in the two feature maps as the output, and the variables are calculated by Eqs. (19) and (20):

$$y^{mean} = f^{mean}\left(x^a, x^b\right) \tag{19}$$

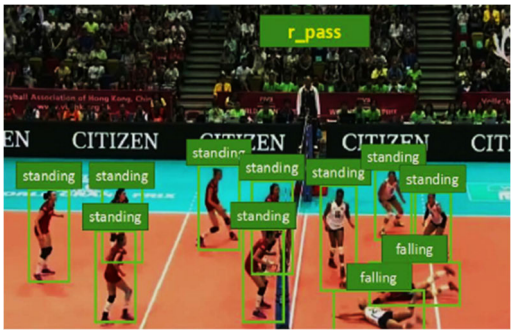

FIGURE 8. The Volleyball dataset image example.

$$y_{i,j,d}^{mean} = mean\left\{x_{i,j,d}^{a}, x_{i,j,d}^{b}\right\} \tag{20}$$

C. EXPERIMENTAL DESIGN

This section installs the Pytorch Deep Learning framework on the 64-bit Ubuntu 16.04 system. The computer Graphic Processing Units (GPU) consist of two NVIDIA GeForce GTX 1080 and one NVIDIA GeForce TITAN XP. There are four Intel Core i7-8700k Central Processing Units (CPUs). The Random Access Memory size is 48G, and Python 3.6 environment is selected as the programming environment. Further, the Adam optimizer is selected as the optimization algorithm of the network model, and the Dropout parameter value is set to 0.5 to prevent overfitting. The initial learning rate is set to 0.001, and the learning rate is set to decay to 0.75 from the previous cycle. This is because the traditional gradient descent strategy will lead to the continuous growth of losses, and too fast a gradient update is easier to overfit. The batch data size is 128. That is, 128 video sequences are processed in each network cycle. The training cycle is 340 cycles. That is, the network conducts 340 training for the entire dataset.

1) DATASET

The research field of this paper belongs to behavior classification. Thus, it does not need to consider specific start and end times but only to classify the actions expressed in the preprocessed and segmented video clips. The dataset used has segmented different categories of action videos. There is no limit to human behavior between categories. Each video has a unique tag or multiple tags. After eliminating the influence of many external factors, the task is a multi-classification problem with action labels. Moreover, the video classification problem is similar to the image processing classification problem. The difference is that it pays attention to the temporal dimension of video. However, feature extraction and classification of video frames can directly use the image processing principle. This section mainly selects three general public datasets to verify the comprehensive performance of the proposed model, including the Volleyball dataset composed of professional volleyball game videos, the UCF-101 dataset, and the HMDB 51 dataset specially used for behavior recognition in video. Specifically, the Volleyball dataset contains 55 volleyball match videos. The producer has marked the corresponding actions of the individual player, including 4,831 frames with action tags, 3,494 training clips, and 1,337 test clips. The tags are divided into individual and group behavior tags. Individual behavior tags include waiting, setting, diving, falling, spiking, blocking, jumping, moving, and standing. Group behavior tags are right set, right spike, right pass, right winpoint, left winpoint, left pass, left spike, and left set. The tags define the group behavior in each part of the game. Each athlete in each group has a bounding box coordinate group and an individual action label annotation to describe their position and individual action. A Volleyball dataset image is depicted in Fig. 8.

The videos in the UCF-101 dataset and HMDB51 dataset come from YouTube and other video websites and public databases. The UCF 101 dataset contains 13,320 videos and 101 behavior categories. The average duration of the video is 7s, and the resolution of each frame is 320*240. It can be divided into five types: human-object interaction, human motion, human interaction, playing musical instruments, and sports. The samples in the HMDB51 dataset mainly come from movie clips. A large number of complex backgrounds and fast-moving shots increase the difficulty of video analysis. The dataset contains 51 human behavior categories and 6,766 videos, with an average duration of 3s. HMDB51 dataset comes from YouTube and Google videos collected over the Internet. Overall, there are 6,849 videos, including 51 categories of 320 $\times$ 240-pixel video frames, with an average duration of 3.0s. There are five types of movements: (1) General facial movements; (2) Facial movements with object manipulation; (3) Whole-body movement; (4) Body movements that interact with objects; (5) Body movements caused by human interaction. Compared with the UCF-101 Dataset, the HMDB51 dataset comes from real scene video, and the background changes greatly. Due to the small amount of data, the network's training is limited. The recognition accuracy of the existing algorithms is generally not high, which has become one of the most challenging datasets. Here, the three datasets' samples are divided by a ratio of 7:3, 70% are used as the training set, and the remaining 30% are used as the test set to verify the model.

2) SPECIFIC EXPERIMENTAL METHODS

The resolution of the video image in the dataset is uniformly adjusted to 224*224. After input into the model, the Faster R-CNN is used to detect and locate the characters in the video, extract the appearance and motion information corresponding to the target, and transmit it to the MS-3DCNN module for feature extraction. Afterward, the integrated three output features are sent to the LSTM network to extract cross-time information. Further, the feature representation of individual behavior is classified by the Softmax function [27], and the behavior recognition result is output. In the experimental process, this paper takes the traditional TSCNN and 3DCNN as the control, and the model on behavior recognition

TABLE 1. Ablation experimental results.

Experimental method	Accuracy (%)
Traditional sequential split two-stream	71.2
3Dres Net101 after expansion	68.8
Traditional timing division two-stream+LSTM	73.7
Expanded 3Dres Net101+LSTM	70.8
The proposed method	75.4

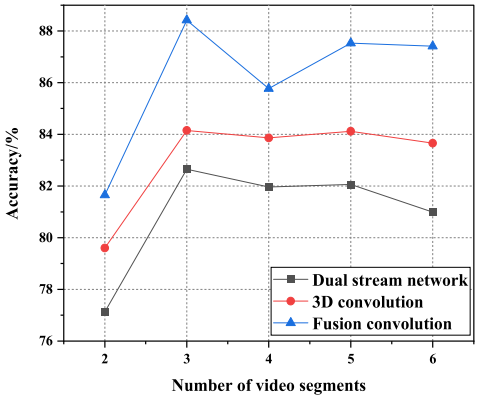

FIGURE 9. Comparison of recognition accuracy of the model under different video segments.

is verified by *Accuracy* index. *Accuracy* is calculated by Eq. (21):

$$Accuracy = \left(\sum_{j=1}^{N} n_{jj}\right) / \left(\sum_{i=1}^{N} \sum_{j=1}^{N} n_{ij}\right) \quad (21)$$

In Eq. (21), n_{ij} refers to the number of samples marked as i and recognized j. n_{jj} is the number of samples marked and recognized as j. N represents the total number of samples. The higher *Accuracy* is, the better the HBR performance of the model for videos.

The number of video segments, the feature fusion method, the fusion proportion of spatial and temporal feature maps, and the learning rate of the model will affect the model's accuracy. Therefore, this section optimizes the above factors based on the Volleyball dataset before the formal experiment. Then, the range of video segments is set to (2, 6). The proportion of the spatial feature map α and time feature map β is set as 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, and 8:2. The learning rate range is set to (0.001, 0.002), and the training iteration 400.

Next, modular progressive ablation experiments are carried out on the Volleyball dataset. The experimental method usually selects the first five frames, the last four frames of the labeled frame, and ten sequential continuous images, including the labeled frame as the input. The number of sequential continuous images, including labeled frames, will be adjusted in the ablation experiment, such as 25 sequential continuous images. The CNN adopts the 3D expansion version of the ResNet and DenseNet, which performs well without pre-training and saves huge pre-training overhead. The resolution of the input image shall be uniformly adjusted to 224 * 224

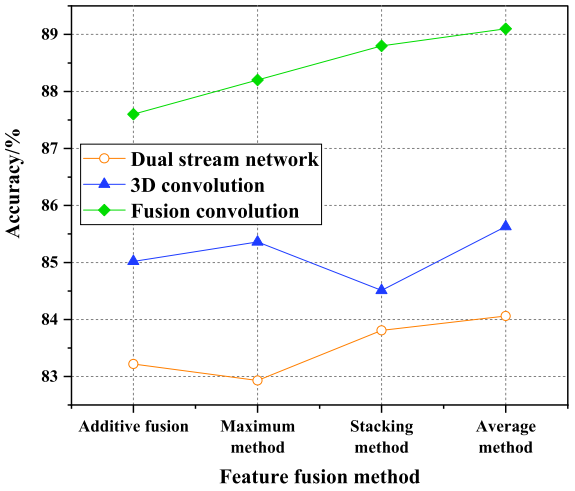

FIGURE 10. Comparison of HBR accuracy under different feature fusion methods.

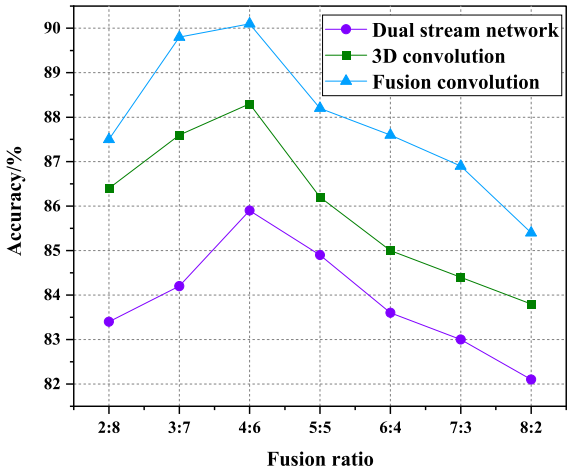

FIGURE 11. Comparison of HBR accuracy of the model under different proportions of Spatio-temporal feature maps.

and processed by data expansion. The data expansion method used in the experiment is multi-scale random cutting, cutting the area defined by the product of minimum length and scale. The proportion is randomly selected from 1.0, 0.875, 0.75, and 0.66. At the same time, the probability of performing the horizontal flip operation is 50% for every three image frames. Then, the appearance features and motion features of the cropped video frames are extracted respectively to meet the needs of the input of the timing segmentation part. The LSTM part adopts a single-layer LSTM network. The input feature vector is 4,096 dimensions, and the LSTM hidden units are 512. The experimental method uses Faster R-CNN as the target detection method, extracts the appearance and motion information of individual targets in the detected scene, and sends them to the spatial-temporal flow networks. The extraction and fusion operations on the 3D convolution layer in the spatial-stream and temporal-stream CNNs are expanded. The output features enter the LSTM

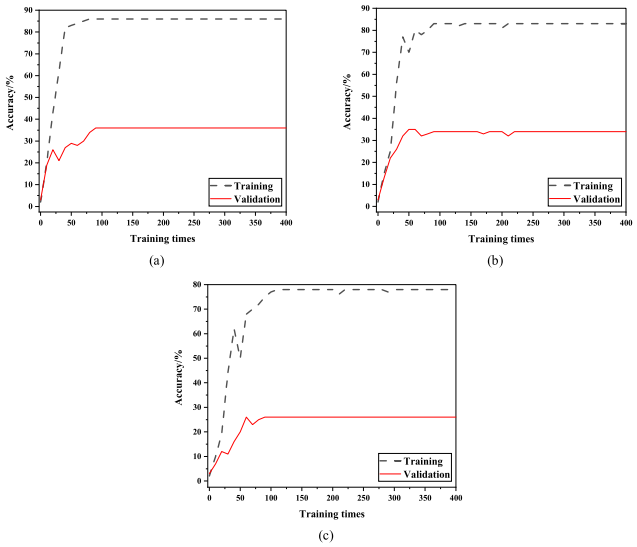

FIGURE 12. Training results of the proposed fusion-CNN-based HBR model on three different training sets (a. Volleyball dataset; b. UCF101 dataset; C. HMDB51 dataset).

network to extract the cross-time information through the connection operation. Based on this, the feature expression of individual behavior is obtained to predict the behavior through the softmax layer classification. The experimental results of real location annotation are used for comparison.

IV. RESULTS OF THE MODEL TEST

The result section completes the proposed model's performance evaluation by comparing it with other models. It includes comparing different video segmentation values and fusion methods, model training and parameter optimization, and behavior recognition tests.

A. ABLATION EXPERIMENTAL RESULTS

The results of the ablation experiment are listed in Table 1:

In Table 1, the two-stream network has higher recognition accuracy than the expanded 3D convolutional network. After adding the LSTM layer, the recognition accuracy of both methods increased. However, the recognition accuracy of the two-stream network method is still higher than the expanded 3D convolution method. The proposed method combines the characteristics of these networks, and its recognition accuracy is higher than each network.

B. COMPARISON OF DIFFERENT VIDEO SEGMENTATION AND FUSION METHODS

1) COMPARISON OF DIFFERENT VIDEO SEGMENTATION

Fig. 9 compares the HBR accuracy of TSCNN, 3DCNN, and the proposed fusion-CNN-based HBR model under different video segments on the Volleyball dataset.

The ordinate of Fig. 9 shows the model accuracy on the test set. As in Fig. 8, when the number of video segments

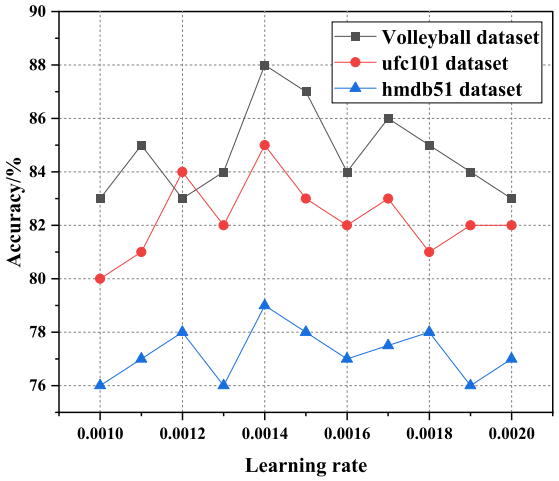

FIGURE 13. HBR accuracy of the proposed model under different learning rates.

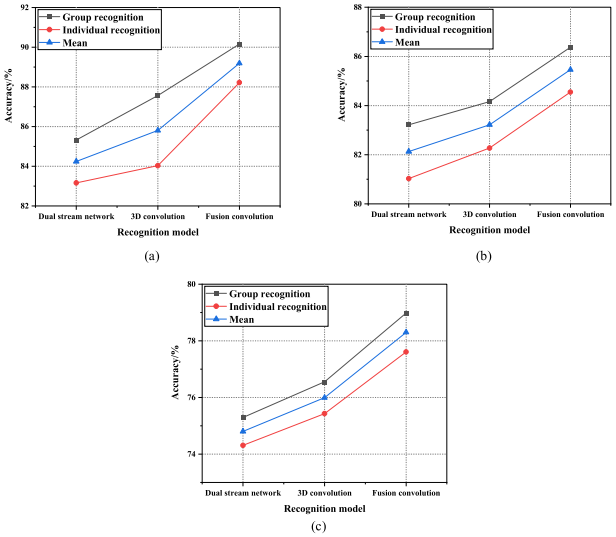

FIGURE 14. Comparison of HBR performance of three models on three datasets (a. Volleyball dataset; b. UCF101 dataset; c. HMDB51 dataset).

is 3, the accuracy of TSCNN, 3DCNN, and the proposed fusion-CNN-based HBR model peaks at 82.65%, 84.15%, and 88.42%, respectively. Moreover, under the same number of segments, the accuracy of the proposed fusion-CNN-based HBR model is higher than the other two models. Therefore, this paper determines to divide the video of the input model into three frame segments. Apparently, when the segmented value is too small and the sample information is insufficient, it results in too simple network training. On the contrary, when the segment value is too large, it results in a large amount of information redundancy and complex calculation, also leading to a decline in network performance. The experimental comparison suggests that the network performance is the best when the segment value is 3. Therefore, this paper divides the video into three segments [28].

2) COMPARISON OF DIFFERENT NETWORK FUSION METHODS

Fig. 10 compares the accuracy of the proposed fusion-CNN-based HBR model under different feature fusion methods: the additive fusion method, maximum value method, stacking method, and average value method.

The ordinate of Fig. 10 is the model accuracy on the test set. As in Fig. 9, the TSCNN, 3DCNN, and the proposed fusion-CNN-based HBR model reach the highest accuracy under the average value method for feature fusion, 84.06%, 85.63%, and 89.1%, respectively. Apparently, the accuracy of the proposed fusion-CNN-based HBR model is higher than that of other models. Therefore, this paper selects the average fusion method to fuse the model output features.

3) COMPARISON OF FUSION PROPORTIONS OF DIFFERENT SPATIO-TEMPORAL FEATURE MAPS

Fig. 11 analyzes the HBR accuracy of the three models under different fusion proportions of the output spatial and temporal feature maps.

The ordinate of Fig. 11 indicates the model accuracy on the test set. As in Fig. 10, the TSCNN, 3DCNN, and the proposed fusion-CNN-based HBR model reach the highest HBR accuracy when the spatial feature map: temporal feature map = 4:6, which are 85.9%, 88.3%, and 90.1% respectively. Similarly, the overall HBR accuracy of the proposed fusion-CNN-based HBR model is higher than that of the other two models. Seven different proportions of weight are taken respectively. When the proportion of spatial feature maps is large, the recognition accuracy decreases; On the contrary, the accuracy increases when the temporal feature map accounts for a large proportion. Hence, the temporal information extracted by the time domain network plays a vital role in the overall network performance. Therefore, the fusion ratio of the spatial and temporal features maps is 4:6.

C. RAINING AND PARAMETER OPTIMIZATION RESULTS

1) MODEL TRAINING RESULTS

Fig. 12 plots the training results of the proposed fusion CNN-based HBR model on three datasets.

The ordinate of Fig. 12 shows the model accuracy on the training set. As in Fig. 11, when the proposed fusion-CNN-based HBR model is trained on the Volleyball dataset, the HBR accuracy reaches the highest (86%) at the 80^{th} iteration. Under the UCF101 dataset, the HBR accuracy reaches the highest (83%) at the 90^{th} iteration. Under the HMDB51 dataset with more complex video features, the HBR accuracy reaches the highest (78%) at the 110^{th} iteration. Overall, the training speed of the proposed model is fast, the accuracy is high, and the performance is excellent. The verification set accuracy is about 30%, and there is a huge gap in training and testing accuracy. The difference between training and verification performance shows insufficient model fitting. Thus, different hyperparametric search techniques need to be used for appropriate training.

2) OPTIMIZATION OF MODEL LEARNING RATE

Fig. 13 draws the results of the proposed fusion-CNN-based HBR model on three datasets under different learning rates using the test set data.

The ordinate of Fig.13 denotes the model accuracy on the test set. The learning rate can control the parameter update speed during model training and impact neural network performance. For example, when the learning rate is too large, the parameters fluctuate near the minimum and cannot approach the optimal solution. By contrast, the model iteration number will increase when the learning rate is too low, and calculation surges and over-fitting might occur [29]. As in Fig. 13, when the learning rate is 0.0014, the HBR accuracy of the proposed model reaches the maximum. Accordingly, the learning rate is set to 0.0014.

D. PERFORMANCE TEST OF THE PROPOSED FUSION-CNN-BASED HBR MODEL

Fig. 14 compares the accuracy of individual HBR and group HBR for TSCNN, 3DCNN, and the proposed fusion-CNN-based HBR model on the three datasets, using the test set data.

The ordinate of Fig. 14 represents the model accuracy on the test set. According to the overall curve trend in Fig. 13, the HBR accuracy of the proposed model is the highest of the three datasets. TSCNN's average HBR accuracy on three datasets is 84.24%, 82.13%, and 74.8%, respectively. By comparison, 3DCNN's average recognition accuracy on three datasets is 85.8%, 83.22%, and 75.99%, respectively. Lastly, the proposed fusion-CNN-based HBR model reaches an average HBR accuracy of 89.89%, 85.46%, and 78.3% on three different datasets.

To summarize, the average accuracy of the proposed fusion-CNN-based HBR model on the Volleyball dataset, UCF101 dataset, and HMDB51 dataset is 4%, 2.7%, and 3% higher than that of TSCNN and 3DCNN, respectively. The proposed model has been significantly improved and is suitable for studying Human Behavior Analysis (HBA) in volleyball match videos. This paper has been applied to the Volleyball dataset and achieved high accuracy in group recognition. Currently, there are various studies on combining CNN with LSTM algorithms for prediction tasks. Similar literature, some scholars conducted a Corona Virus Disease 2019 (COVID-19) hotspot prediction based on a deep hybrid neural network. The model prediction accuracy by combining CNN and LSTM is 78%. By comparison, the prediction accuracy of this paper is slightly higher than the literature value.

V. CONCLUSION

This paper constructs an HBR model based on fusion-CNN and makes up for the shortcomings of CNN and LSTM. The experimental design helps optimize the model's key parameters by selecting the open data set. The experimental results show that the algorithm parameter settings greatly impact the model's recognition accuracy, and the recognition accuracy

of the proposed fusion-CNN model is the highest. The average accuracy of HBR of the proposed fusion-CNN on the Volleyball dataset, UFC101 dataset, and HMDB51 dataset is 4%, 2.7%, and 3% higher than the TSCNN and 3DCNN, respectively. The model improvement effect is significant and suitable in Volleyball game video behavior research. Compared with other similar studies, this paper has a higher prediction accuracy. The shortcomings of this paper are noticeable. The recognition accuracy for individual behavior needs to be improved. It only explores the performance of individual behavior and group behavior recognition separately, ignoring the possible correlation between individual behavior and group behavior recognition. Because of the use of classic datasets, this paper defaults that the datasets are of high quality. No evaluation indicators for the quality of datasets are set. The model is not comprehensively evaluated with new datasets. There is a big gap between the training and testing accuracy in the research, indicating that the model is overfitted. The research does not provide an adaptive learning method. The focus is on researching players' volleyball data sets, neglecting players' postures and behavior. There are few evaluation and measurement methods for model performance. In future research, the accuracy of individual behavior recognition of the model will be improved, the mechanism of individual behavior affecting the accuracy of group behavior recognition will be discussed, and a more comprehensive performance evaluation of the model will be conducted in combination with new data sets. It is expected to analyze new adaptive learning methods to solve overfitting and explore more cost-effective research methods based on player posture and behavior. Additionally, more metrics will be considered to evaluate the model performance comprehensively. The behavior recognition model based on fusion-CNN proposed can provide more accurate results for volleyball video behavior recognition, which is of great significance for promoting the rapid development of volleyball events.

REFERENCES

[1] W. Gu, J. Zhang, J. Zhang, Z. Li, and L. Nie, "Study on the influence of volleyball sports on the construction of physical culture in colleges and universities," *World Sci. Res. J*, vol. 7, pp. 47–51, Jan. 2021.
[2] K. Aini, M. Asmawi, and R. Pelana, "Games based model of volleyball passing exercise for junior high school student," *ACTIVE, J Phys. Educ., Sport, Health Recreat.*, vol. 9, pp. 17–22, Mar. 2020.
[3] F. A. Salim, F. Haider, D. Postma, R. van Delden, D. Reidsma, S. Luz, and B.-J. van Beijnum, "Towards automatic modeling of volleyball players' behavior for analysis, feedback, and hybrid training," *J. Meas. Phys. Behav.*, vol. 3, no. 4, pp. 323–330, May 2020.
[4] J. Chang, S. Hong, D. Son, H. Yoo, and H. Ahn, "Development of real-time video surveillance system using the intelligent behavior recognition technique," *J. Inst. Internet, Broadcast. Commun.*, vol. 19, no. 2, pp. 161–168, Feb. 2019.
[5] Y. Chen, S. Hu, H. Mao, W. Deng, and X. Gao, "Application of the best evacuation model of deep learning in the design of public structures," *Image Vis. Comput.*, vol. 102, Oct. 2020, Art. no. 103975.
[6] M. Hu, Y. Zhong, S. Xie, H. Lv, and Z. Lv, "Fuzzy system based medical image processing for brain disease prediction," *Frontiers Neurosci.*, vol. 15, p. 965, Jul. 2021.
[7] C. Liu, J. Ying, H. Yang, X. Hu, and J. Liu, "Improved human action recognition approach based on two-stream convolutional neural network model," *Vis. Comput.*, vol. 37, no. 5, pp. 1327–1341, Jun. 2021.
[8] Y. Hu, "Design and implementation of abnormal behavior detection based on deep intelligent analysis algorithms in massive video surveillance," *J. Grid Comput.*, vol. 18, pp. 227–237, Aug. 2020.
[9] J.-C. Chen, C.-Y. Lee, P.-Y. Huang, and C.-R. Lin, "Driver behavior analysis via two-stream deep convolutional neural network," *Appl. Sci.*, vol. 10, no. 6, p. 1908, Mar. 2020.
[10] L. Chen, R. Liu, D. Zhou, X. Yang, and Q. Zhang, "Fused behavior recognition model based on attention mechanism," *Vis. Comput. Ind., Biomed., Art*, vol. 3, no. 1, pp. 1–10, Sep. 2020.
[11] S. Lan, S. Jiang, and G. Li, "An elevator passenger behavior recognition method based on two-stream convolution neural network," *J. Phys., Conf.*, vol. 1955, no. 1, Jun. 2021, Art. no. 012089.
[12] Z. Du, G. Zhang, W. Lu, T. Zhao, and P. Wu, "Spatio–temporal transformer for online video understanding," *J. Phys., Conf.*, vol. 2171, May 2022, Art. no. 012020.
[13] A. Dhillon and G. K. Verma, "Convolutional neural network: A review of models, methodologies and applications to object detection," *Prog. Artif. Intell.*, vol. 9, no. 2, pp. 85–112, 2020.
[14] R. P. de Lima and K. Marfurt, "Convolutional neural network for remote-sensing scene classification: Transfer learning analysis," *Remote Sens.*, vol. 12, no. 1, p. 86, Sep. 2020.
[15] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, pp. 611–629, Sep. 2020.
[16] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Mar. 2021.
[17] L. Wang, L. Wang, H. Lu, P. Zhang, and R. Xiang, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, Apr. 2018.
[18] Y. Yu, X. Si, C. Hu, and Z. Jianxun, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Nov. 2019.
[19] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *Eur. J. Oper. Res.*, vol. 270, no. 2, pp. 654–669, Jul. 2018.
[20] Le, Ho, Lee, and Jung, "Application of long short-term memory (LSTM) neural network for flood forecasting," *Water*, vol. 11, no. 7, p. 1387, Apr. 2019.
[21] E. Zhang, B. Xue, F. Cao, J. Duan, G. Lin, and Y. Lei, "Fusion of 2D CNN and 3D DenseNet for dynamic gesture recognition," *Electronics*, vol. 8, no. 12, p. 1511, Dec. 2019.
[22] W. Zhou, Z. Chen, and W. Li, "Dual-stream interactive networks for no-reference stereoscopic image quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3946–3958, Oct. 2019.
[23] T. Akilan, Q. J. Wu, A. Safaei, J. Huo, and Y. Yang, "A 3D CNN-LSTM-based image-to-image foreground segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 959–971, Aug. 2019.
[24] Y. Ren, C. Zhu, and S. Xiao, "Small object detection in optical remote sensing images via modified faster R-CNN," *Appl. Sci.*, vol. 8, no. 5, p. 813, Dec. 2018.
[25] L. Zhang and H. Schaeffer, "Forward stability of ResNet and its variants," *J. Math. Imag. Vis.*, vol. 62, no. 3, pp. 328–351, Jan. 2020.
[26] S.-L. Lin, "Application combining VMD and ResNet101 in intelligent diagnosis of motor faults," *Sensors*, vol. 21, no. 18, p. 6065, Feb. 2021.
[27] Q. Zhu, Z. He, T. Zhang, and W. Cui, "Improving classification performance of softmax loss function based on scalable batch-normalization," *Appl. Sci.*, vol. 10, no. 8, p. 2950, Jun. 2020.
[28] C. Zhang, E. Nateghinia, L. F. Miranda-Moreno, and L. Sun, "Winter road surface condition classification using convolutional neural network (CNN): Visible light and thermal image fusion," *Can. J. Civil Eng.*, vol. 49, no. 4, pp. 569–578, May 2022.
[29] V. Bijalwan, V. B. Semwal, G. Singh, and T. K. Mandal, "HDL-PSR: Modelling spatio–temporal features using hybrid deep learning approach for post-stroke rehabilitation," *Neural Process. Lett.*, vol. 16, pp. 1–20, Jan. 2022.

CHEN LIANG received the bachelor's degree in sports training in 2016 and the master's degree in physical education and training from Beijing Sport University, in 2020, where he is currently pursuing the Ph.D. degree with the China Volleyball College. He had participated in the publication of three books and ten papers and a number of subjects, including volleyball artificial intelligence service machine and children's volleyball and carried out research in sports training, artificial intelligence, and image processing. His research interests include the principle and method of physical education teaching and training. He has been awarded the First Level National Referee of Volleyball in China and has served for many matches. He has also published two research papers at prestigious international conferences and journals. During his tenure, he improved the classroom learning environment and academic and professional curricula to accommodate improved athletic training influenced by technology.

ZHIJUN LIANG received the master's degree in physical education and training from Beijing Sport University, in 2003. He is currently a Teacher with the China Volleyball College, Beijing Sport University, where he used to be the Dean Assistant and the Head of the Teaching Team. Be qualified as the National Referee of Volleyball, International Referee of Sitting Volleyball. His part-time job includes a Statistics Training Lecturer at China Volleyball Super League, a member of the Referee Committee of China Paravolley Federation, the General Secretary of Paravolley Asian-Oceania (PVAO), Level-3 International Technical Official of World Paravolley (WPV), and a Training Lecturer at the International Technical Officials of World Paravolley. He is also a full-time Teacher with the China Volleyball College. From 2007 to 2008, be employed at the Technology Department of Beijing 2008 Olympic Organizing Committee as a Level-4 Event Expert, responsible for the statistics and results system of volleyball and sitting volleyball. He has presided over one universal-level project, participated in eight provincial-level projects, completed a monograph independently, participated in the compilation of three textbooks, completed one technical report of international individual organizations, participated in an international conference, and completed an oral report.