IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 December 2023, accepted 27 December 2023, date of publication 1 January 2024,
date of current version 16 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3349020

RESEARCH ARTICLE

The Computer Vision Simulation of Athlete's Wrong Actions Recognition Model Based on Artificial Intelligence

WENXIN DU
School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
e-mail: mnkmYuki@sjtu.edu.cn

This work was supported by the National Key Research and Development Program of China under Grant 2021ZD0110704.

ABSTRACT At present, in basketball teaching in China, the traditional basketball training method is for coaches to communicate face-to-face with athletes, observe their basketball movements, and judge the correctness of the movements based on the coach's personal experience. However, this method mainly relies on the subjective judgment of the coach and lacks objective evaluation of athletes, making it impossible to objectively evaluate their performance. This article mainly studies an athlete's incorrect action recognition model based on artificial intelligence algorithms and computer vision, and constructs an athlete's incorrect action recognition model based on a dual channel 3D convolutional neural network (CNN). In this project, spatial attention mechanism (SA) was introduced into 3D CNN. By using inter frame difference information that can represent significant changes in athletes' motion status, and combining it with grayscale video data, accurate recognition of athletes' incorrect actions was achieved. The simulation results show that as the number of basketball technical errors increases, the recognition accuracy of this method decreases slowly. When the number of basketball technical errors reaches 400, the accuracy of action recognition is still as high as 87.552%. This indicates that this method can control the error rate within a reasonable range, improve the ability to identify basketball technical errors, and provide strong support for basketball teaching. In addition, the experimental results of this method also include various other achievements in performance calculation, further verifying its superiority in identifying basketball technical errors.

INDEX TERMS Computer vision, artificial intelligence, athlete, wrong actions, action recognition, 3D CNN.

I. INTRODUCTION

Human motion and motion recognition is an important subject in CV(Computer vision). In recent years, with the popularization of CV technology in China, image data and image data have been applied to the analysis of human body structure, and they can show different postures when moving. In basketball, the correct posture is conducive to the improvement of players' technical level, while the improper posture will have an adverse impact on players' technical level. Therefore, it is very important to identify and correct the posture of basketball players accurately for them to improve their skills and achieve good competition results. In daily training, athletes' needs for quality evaluation of movement completion and improvement of sports effect cannot be fully met. Therefore, after AI(artificial intelligence) develops rapidly and matures gradually, it is very necessary to study the algorithm of simulating human eyes' judgment based on the recognition and evaluation of gymnasts' movements from the perspective of deep learning, aiming at the typical characteristics of gymnastics, such as large amplitude and high speed.

At present, human motion analysis is one of the most active research topics in CV field, and its core is to detect, track and identify people from image sequences by CV technology, and to understand and describe their behaviors [1]. It is

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

proposed to use 3D bone nodes in depth images to restore human movements and realize human movement recognition [23], [13] used CNN (Convolutional Neural Network) and DT(Decision tree) to extract and classify the acceleration data obtained by the acceleration sensor built in the smart phone. Avola et al. [2] use dynamic module matching algorithm to detect and identify the wrong posture of basketball players and complete the posture identification and correction of basketball players; The dual-stream network proposed by Zhang et al. [30], G and others adds optical flow information to the ordinary RGB video stream to improve the recognition effect, but it does not make full use of the video information including depth and bone information. At the same time, the calculation of optical flow also increases the time complexity, which makes it difficult to recognize online video. Xu and Luo [27] put forward an image reconstruction method which uses the characteristic shift of differential scanning to move the error between frames. This method can effectively improve the error, but there are some problems such as low error rate. Nunez-Marcos et al. [19] established a model based on LSTM (Long Short-Term Memory). This network established two sub-networks by using attention mechanism, and paid different degrees of attention to different frames and joints, and finally realized end-to-end behavior recognition.

In the aspect of identifying athletes' wrong actions, 3D visual detection modeling technology mainly discriminates athletes' wrong actions by 3D vision, and then detects and evaluates their wrong actions by 3D vision [14]. This paper mainly studies the model of athlete's wrong action recognition based on AI algorithm and CV, constructs the model of athlete's wrong action recognition based on dual-channel 3D CNN, expands the convolution kernel dimension, and performs convolution operation in two dimensions of time and space to extract human action features. Significant areas in human motion are represented by inter-frame difference, and dual channels are constructed by using frame difference and original gray video frames to assist 3D CNN model in human motion classification and recognition. There is a significant gap between this study and existing literature in the field of athlete error recognition. Firstly, this study has made significant breakthroughs in action recognition in complex backgrounds. Traditional methods often find it difficult to accurately recognize athletes' movements in complex and chaotic backgrounds, as obstacles and others may obstruct the athletes' bodies, affecting the accuracy of the model. In contrast, this study was able to better cope with recognition challenges in complex backgrounds and improve the accuracy of erroneous actions by introducing new technologies and algorithms. Secondly, the study focuses on the adaptability of action scales. In real-life scenarios, the scale of athletes' actions may vary depending on the camera distance, and traditional methods are difficult to handle this situation, which can easily lead to feature extraction errors. This study introduces algorithms for scale changes to ensure accurate feature extraction at different distances, thereby improving the robustness of action recognition. Thirdly, this study utilizes advanced computer vision technologies, including 3D vision, which provides richer information and depth compared to traditional 2D image analysis methods. The introduction of this technology makes the recognition of incorrect actions more contextual and contextual, thereby improving accuracy. Finally, the study also utilized feature descriptor prediction, combining information observed from the previous and current frames to accurately encode human motion information. This innovative method not only reduces the dependence on precise model construction, but also improves the efficiency of the algorithm.

This study has made significant progress compared to existing literature through innovation in recognition in complex backgrounds, action scale adaptation, introduction of 3D vision, and feature descriptor prediction, bringing more accurate and efficient solutions to the field of error action recognition. These gaps highlight the uniqueness and foresight of this study, which is expected to make important contributions to the recognition of erroneous actions and the improvement of research level in the field. This work stands out in the field of athlete's wrong action recognition by addressing challenges related to complex backgrounds, scale changes, and temporal positioning, while also incorporating advanced CV techniques and feature descriptor prediction to enhance the accuracy and efficiency of the recognition process. These contributions collectively advance the state of the art in the field and offer promising prospects for improved athlete's wrong action identification.

II. RESEARCH METHOD

A. ACTION FEATURE EXTRACTION

In the realm of athlete's wrong action recognition, this work introduces several novel contributions that set it apart from existing literature. Human movements are inherently non-rigid and pose challenges in expressing and classifying them accurately. One distinguishing aspect of this work is its approach to address the complex and chaotic background scenarios common in sports settings. In such situations, obstacles and other individuals can obstruct the athlete's body, making accurate body localization difficult. Furthermore, varying lighting conditions can impact the athlete's appearance, potentially leading to the failure of conventional fixed human appearance models [3]. Another notable innovation lies in the adaptation to scale changes inherent in athlete's movements. The athlete's wrong action recognition algorithm employed here has been designed to handle the scaling of human bodies, ensuring that features are correctly extracted even when the distance from the camera affects body size. Additionally, this method also focuses on accurately locating the temporal position of human actions in the time domain, contributing to improved recognition accuracy. This work also builds upon existing methods in computer vision (CV) by incorporating advanced techniques. While CV initially centered on 2D image analysis, the research community has increasingly ventured into 3D vision [25]. Leveraging 3D vision capabilities enhances the accuracy and depth of athlete's wrong action recognition, enabling the computer to comprehend the context and meaning of actions within video scenes. Moreover, this work integrates feature descriptor prediction, combining information from previous frames and current frame observations to accurately encode human motion information. This innovative approach reduces reliance on precise model construction while maintaining efficiency by avoiding a time-consuming number of model states [21].

The two most important links are feature extraction and description and classification recognition(Figure 1). These two processes transform complex video data types into features in the form of matrix, vector and numerical value containing motion information, and then use classifiers to identify and classify these feature data [26].

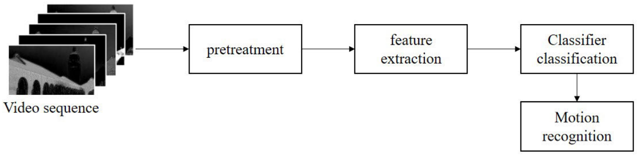

FIGURE 1. The basic process of identifying athletes' wrong actions.

At present, the human pose estimation technology has been developed rapidly, especially the 2D pose estimation technology. However, in the practical application process, due to the influence of the appearance of the object, joint position, background interference, occlusion and other factors, the pose estimation effect of the object is not good In sports, because of the great diversity of athletes' movements, it is not suitable to use 2D posture method. In contrast, the estimation method of 3D human posture is more suitable for practical application. At present, researchers mainly use 3D human posture estimation network based on 2D posture data for semi-supervised learning to improve the accuracy of the network model.

Future work directions include further improving the accuracy of action recognition in complex scenarios. Although this study has made significant progress in complex contexts, challenges still exist in real-world competition and exercise scenarios, such as interactions between multiple athletes and different lighting conditions. Future research can explore more advanced background modeling and lighting correction techniques to further improve recognition performance [22]. In addition, for the issue of scale changes, future work can consider applications in more contexts. Different sports events and venues may lead to broader scale changes, requiring more flexible and intelligent scale adaptation algorithms [28]. However, this study also has some limitations. Firstly, despite the introduction of advanced technologies and algorithms, performance degradation may still occur under extreme conditions such as strong light, low light, or athlete occlusion [13]. Future work needs to further improve the robustness of the algorithm. In addition, the experimental dataset of this study may be limited, and future work can consider more diverse and realistic datasets to verify the robustness of the algorithm. Finally, the method of this study may require high hardware resources, and future work can explore more effective computational methods to reduce costs.

Sports activities are divided into individual events and group events, and all kinds of sports videos are analyzed. In team events, we will focus on the strategies adopted by the two teams and evaluate the overall performance of the teams. In individual events, it is mainly to analyze the athletes' posture and movement accuracy, so as to help athletes train and evaluate their performance. In 3D human posture estimation, there are many methods based on a single image. Since the 2D pose can correspond to the projection of multiple 3D poses, the network adjustment uses the images projected by the whole body 3D mesh and 2D annotations to optimize the projection consistency and reduce the dependence on the real value of 3D joints. Upgrading from 2D human pose to 3D human pose will create uncertainty in 3D space, and the acquisition of depth information can reduce the uncertainty.

In the process of feature extraction of athletes' action images, the athletes' actions are expressed as a continuous state sequence, and each state has its own apparent and dynamic features. On this basis, the local features of athletes' action images are extracted, and the spatial and temporal neighborhoods of all the local features of athletes' action images are divided into multi-scale grids, and the distribution of local features of athletes' action images is counted. The sports technology is connected to form the overall features of athletes' action images, and the feature extraction of athletes' action images is completed [16].

The comparative analysis of performance parameters of existing technical methods shows the advantages and disadvantages of different methods in athlete error motion recognition. Firstly, the model based on deep learning performs well in complex action recognition and has high accuracy. These models can automatically learn features and capture spatial and temporal information of actions, thus possessing strong robustness in complex backgrounds and different scales. However, these models typically require a large amount of labeled data and computational resources, and the complexity of the models is high, which may require longer training time. In contrast, some traditional methods perform better in computational efficiency, but there are certain limitations in accuracy and robustness. These methods typically rely on manually designed feature extractors and may struggle to adapt to complex actions and background changes. In addition, they may be more sensitive to light and occlusion. Overall, the comparative analysis of performance parameters indicates that deep learning based methods have advantages in accuracy and robustness, but require more data and computational resources. Traditional methods have advantages in computational efficiency, but may not perform well in complex scenarios. Therefore, in practical applications, the selection of appropriate methods should be balanced based on specific scenarios and available resources to achieve the best performance in athlete error motion recognition. Future research may focus on finding a better balance between accuracy and efficiency, and improving the robustness of the model.

Because the competition place of athletes is in a limited area, athletes will exercise in such a space. Through the establishment of 3D coordinate system, the athletes' movement states can be clearly understood in the process of feature extraction. Then a 3D coordinate system as shown in Figure 2 can be constructed.

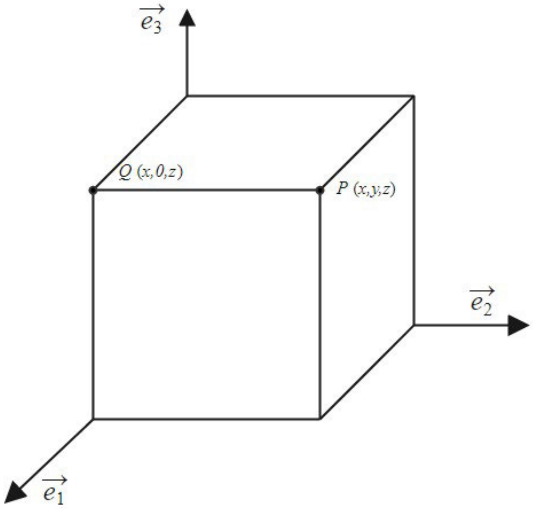

FIGURE 2. **3D coordinate system.**

In the process of identifying and correcting the athlete's posture, $Q = (Q_G, Q_G, Q_B)$ represents the foreground pixel of the athlete's posture image and $W = (E_G, E_G, E_B)$ represents the background pixel of the athlete's potential image, and each joint point of the athlete's basketball is located by Formula (1):

$$G(x, y, z) = \frac{Q * W}{(\Delta_T, \Delta_D, \Delta_0, \Delta_{L1}, \Delta_{R0}, \Delta_{R1})} \tag{1}$$

According to the data field, corner marking and 3D template matching are carried out on the wrong action image sequence in the moving state in sports, so as to realize the reconstruction of the wrong action image in sports. Construct the difference feature quantity of wrong action images in sports [11], and find out the sparse linear equations for template matching of wrong action images in sports as follows:

$$g(x, y) = h(x, y) \oplus f(x, y) + \eta(x, y) \tag{2}$$

where $h(x, y)$ is the video tracking parallax function of the wrong action image in sports, and symbol $\oplus$ represents convolution.

3D skeleton data is a depiction of human body in real 3D space, which abstracts human body into several joints connected with bones, thus providing a more concise 3D description of human body without losing accuracy. Kinect equipment first uses the depth value in the depth map to segment the background and people, and obtains a depth map containing only human body [6]. Then, according to the pre-trained model for recognizing human body parts, the obtained depth map of human body is segmented, and finally, the information of joint points of the segmented human body parts is marked.

Most of the local features used to identify athletes' wrong actions are derived from local feature description operators. Unlike the statistical method of HOG (Histogram of Oriented Gradient) features, HOF (Histogram of optical flow) features only count the optical flow histogram [4]. Therefore, the HOF feature can alleviate the problem that the action scale of optical flow feature is too sensitive.

Assuming that $P_t\ (x_t, y_t)$ is the position coordinate of the feature point in the I_t frame, the position coordinate of the feature point in the next frame can be obtained by Formula (3).

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)\big|_{x_t, y_t} \tag{3}$$

where M stands for a median filter with a size of 3∗3, $\omega_t = (u_t, v_t)$ stands for the optical flow information calculated from the I_t frame and the I_{t+1} frame, and u_t, v_t stands for the optical flow generated in the horizontal and vertical directions. The motion direction of a feature point can be calculated from the median value of optical flow in the neighborhood of the point.

In the process of feature extraction of athletes' action images, assuming that (x, y, t, σ, τ) represents a continuous sequence of athletes' action states and (x, y, t) represents the dynamic features of athletes' actions, the overall feature extraction model of athletes' action images obtained by connecting the distribution of multiple grids is established by using Formula (3) to complete the feature extraction of athletes' action images:

$$H = \frac{\det(\mu) - ktrace^3(\mu)}{(x, y, t, \sigma, \tau) * (x, y, t)} \tag{4}$$

where μ represents the spatial multi-scale grid division of all local features of motion images, k represents the time neighborhood multi-scale grid division of all local features of motion, and $trace^3$ represents the statistical distribution of local features of motion images formed by motion technology.

B. OPTIMIZATION OF DETECTION METHOD OF ATHLETES' WRONG ACTIONS IN SPORTS

In the training of basketball players, how to accurately identify the movements in difficult videos is a difficult point. Only by accurately identifying the athletes' movements can we provide a better basis for the follow-up training of basketball players. Video motion recognition technology has always been the focus of research at home and abroad, especially after the development of CV technology has become more and more mature, the computer is mainly used to identify video data for processing and recognition [17]. At present, the characteristics of basketball players' foul behavior are mainly expressed by a group of symbols, and their characteristics are determined by decomposing a single symbol. However, this method is difficult to accurately extract the characteristics of foul behavior in high-level basketball games. The computational complexity of the proposed work will be a key consideration. Deep learning based methods typically require a large amount of computational resources, including high-performance computers and GPU acceleration, to train complex neural network models. This may result in longer training time and high hardware costs. At the same time, processing large-scale motion datasets also requires a large amount of storage and computing power. In order to address computational complexity, future work can focus on optimizing algorithms and models to improve computational efficiency. In addition, distributed computing and cloud computing resources can be considered to accelerate the training and inference process. In addition, data sampling and dimensionality reduction techniques can also be used to reduce computational burden. In summary, computational complexity will be a challenge for future research on athlete error recognition, requiring comprehensive consideration of factors such as algorithms, hardware, and data management.

Due to the rapid development of science and technology, especially Internet technology in recent decades, the research on AI has also turned into distributed research, and the research object has also changed from a single intelligent agent to an intelligent group research. Since then, AI has started to be practical. AI is a subject about knowledge, that is, how to express knowledge and how to acquire and use knowledge.

Traditional bone-based behavior recognition algorithms often use a set of artificially constructed features to imitate human behavior. However, there are still some defects in the existing research, such as the inability to fully explore the spatial relationship between people. In recent years, the technology of motion error recognition based on deep learning theory has made some achievements in motion error recognition. In recent years, according to the time series characteristics of human bones, many researchers combined data model with graph structure and proposed a motion prediction method based on CNN. In addition, in the CV field, when information processing encounters bottlenecks, attention mechanism will be introduced into the neural network to help focus some important information, so that the network can obtain more valuable feature information and further improve the network performance [20].

The image classification and recognition technology based on CNN has made great progress in image classification and recognition. Compared with traditional image recognition technology, it has great advantages in recognition accuracy and generalization ability. Compared with 2D CNN, 3D CNN expands the time domain in spatial domain, and can directly extract spatial domain and spatial domain features from RGB images. 3D convolutional neural network can directly extract spatial-temporal-spatial features from RGB images, but it has some shortcomings, such as high model complexity, many parameters, large computation and large storage space, which restricts the depth of the neural network and leads to its inability to obtain richer and more abstract features [10].

This project plans to construct SA(Spatial attention) mechanism in 3D CNN, and combine it with the gray image auxiliary network, and use the gray image auxiliary network to identify the wrong actions of athletes. The dual-channel 3D CNN model constructed in this paper is shown in Figure 3, which includes a dual-channel input layer, five 3D convolution layers and five 3D pooling layers interlaced with each other. Finally, the classification results are obtained after connecting two fully connected layers, and two Dropout are performed on the two fully connected layers respectively.

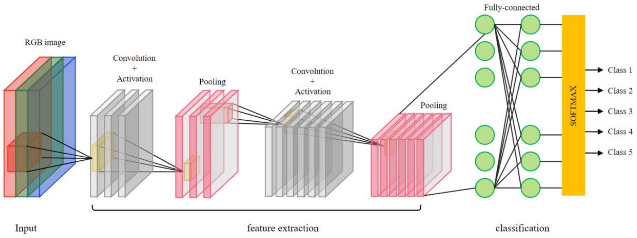

FIGURE 3. Dual-channel 3D CNN structure.

The input video first passes through a 3D convolution layer with a convolution kernel size of 7 * 7 * 7 and a step size of 2 * 2 * 1, and the L2 regularization method is used to initialize the weight of the convolution kernel. The image uses RGB three-channel images that have been preprocessed from the original video.

Before transmitting the bone data to the primitive, the bone data is normalized, thus improving the standardization and operability of the bone data. This action will be carried out at the standard level of batch. After passing through 9 primitives, the feature map is transmitted to the global average gallery, thus obtaining a feature vector with a certain scale. At the end of the network structure, a Softmax classifier is introduced, which can calculate the scores of various behavior types and classify behaviors into behavior types, thus predicting behaviors. The computational complexity of the proposed work will be a key consideration. Deep learning based methods typically require a large amount of computational resources, including high-performance computers and GPU acceleration, to train complex neural network models. This may result in longer training time and high hardware costs. At the same time, processing large-scale motion datasets also requires a large amount of storage and computing power. In order to address computational complexity, future work can focus on optimizing algorithms and models to improve computational efficiency. In addition, distributed computing and cloud computing resources can be considered to accelerate the training and inference process. In addition, data sampling and dimensionality reduction techniques can also be used to reduce computational burden.

In summary, computational complexity will be a challenge for future research on athlete error recognition, requiring comprehensive consideration of factors such as algorithms, hardware, and data management.

On this basis, a 3D CNN model based on SA is proposed to identify athletes' wrong actions. In the detection and segmentation of moving objects, the frame difference method is a common algorithm. The operation flow includes: firstly, subtracting the corresponding pixel values of adjacent frame images to obtain a difference image, and then performing opening and closing operations on the difference image and setting a threshold to binarize it. In the spatial map stack, after the attention module of the map, the preliminary extracted spatial features are obtained. In order to get a better representation of action features, SA mechanism is introduced, and SA module is added after spatial graph convolution. The structure of this module is shown in Figure 4.

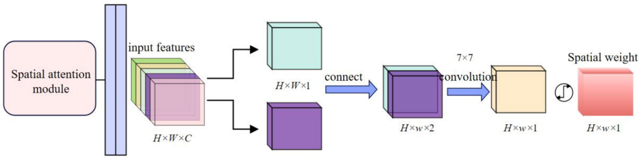

FIGURE 4. SA module.

In the convolution layer, each layer often produces new channels, and each channel has different correlation with key information. Therefore, the signal on each channel can be given corresponding weight, and the output $f_{out} \in R^{H*W*C}$ can be used as the input of the module for squeezing operation to realize global information embedding.

This operation can be expressed by the following formula:

$$z_c = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{i=1}^{W} m_c(i, j) \tag{5}$$

where $m_c \in R^{H*W}$ is an element of the matrix z output through this step. This formula represents the average pooling operation in time and space dimensions.

Next, transform the output z, as shown in the following formula:

$$S = \sigma\left(W_2 \delta\left(W_1 z\right)\right) \tag{6}$$

where $W_1 \in R^{\frac{C}{r} \times C}$, $W_2 \in R^{C \times \frac{C}{r}}$ is the two weight matrices of the fully connected layer, σ stands for Sigmoid activation function and δ stands for ReLu activation function. Multiply S with the input feature map f_{out}.

The 3D convolution formula is as follows:

$$F' = F \oplus K \tag{7}$$

$$F_j^{'xyz} = \sum_{n} \sum_{h=0}^{h_k-1} \sum_{w=0}^{w_k-1} \sum_{t=0}^{t_k-1} K_{jn}^{hwt} F_n^{(x+h)(y+w)(z+t)} \tag{8}$$

where $\oplus$ represents the 3D convolution operation, $F_j^{'xyz}$ represents the value of the j-th feature map at the position of (x, y, z) space, $j = n_k$, that is, the number of feature maps of $F'$is equal to the number of 3D convolution kernels K, and K_{jn}^{hwt} represents the value of the j-th convolution kernel connected with the n-th feature map of F at the space coordinate (x, y, z).

The residual structure has one more short-cut branch than the traditional convolution structure, which is used to transmit the bottom information to the upper network layer, so that the network can be trained deeply. The residual network is dedicated to learning the stagger $H(x) - x$ between input and output by using multiple parameter network layers, namely:

$$x- > (H(x) - x) + x \tag{9}$$

where $H(x) - x$ is the residual between input and output to be learned by these multi-parameter network layers, that is, residual learning, and x is to directly establish an identity mapping between input and output in the residual structure as shown on the right.

Because of the limited capacity of input images, 3D CNN only needs to extract some images from the video as input, and the sampling rules for generating video sample frames have a great influence on the generalization ability of the network model.

In order that each sample can cover more key video frames, this paper uses the frame interval R to get L frame samples. For the i sample, the sampling frame subscript rule is as follows:

$$C_i = \{S_i, S_i + R, S_i + 2R, \cdots, S_i + (L-1)R\} \tag{10}$$

where S_i is that subscript of its start frame.

III. ANALYSIS AND DISCUSSION OF SIMULATION RESULTS

The data of athletes' wrong technical actions collected in basketball matches are adopted. The video sequence used is obtained by framing the video in the basketball basic action data set, and each video sequence is treated as a gray image before the experiment.

All experiments are based on Windows 10.0 system, and the running platform is Nvidia GeForce GTX 1080Ti GPU. PyCharm is used as an integrated development environment, and PyTorch deep learning framework is adopted. In order to quickly determine the number of frames in the sampling interval, this experiment only selects the top 10 sports videos in the basketball basic action data set for experimental analysis, and the data set is divided into 80% training set and 20% test set.

Taking 35 frames as samples and taking 3, 4 and 5 frames as sample intervals, a comparative experiment was conducted. On this basis, through 30 cycles of repeated training, the average correct rate of athletes' identification of wrong actions in sports is calculated. Table 1 provides the test results.

When the sampling interval is set to 5, the test accuracy of the model is 85.78% after 20 rounds of iterative training. For

TABLE 1. Accuracy of different sampling results after 30 rounds of model training.

Sampling frame number	Sampling interval	Recognition accuracy(%)
35	3	75.71
35	4	79.788
35	5	85.78

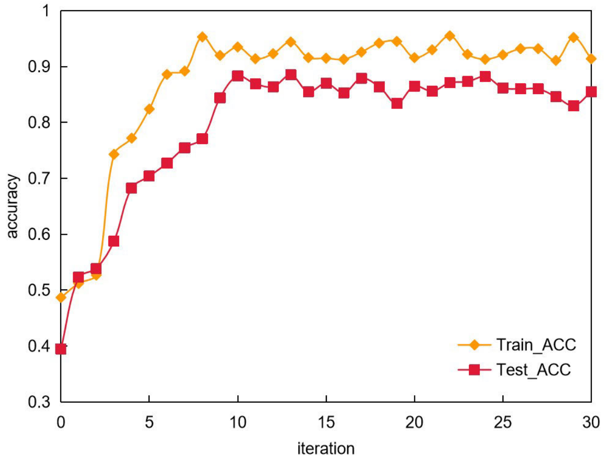

FIGURE 5. Accuracy curve of training set and test set.

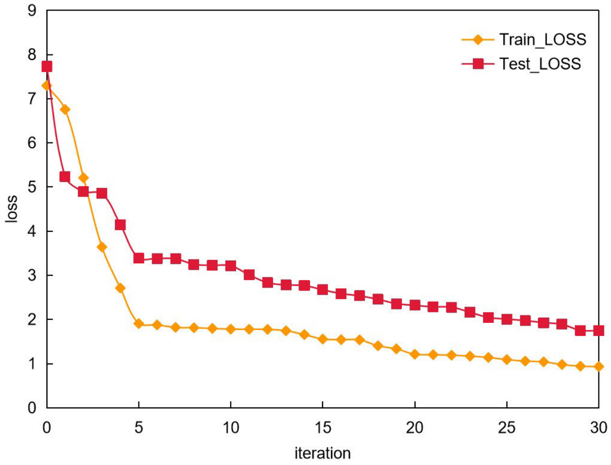

FIGURE 6. Error curve of training set and test set.

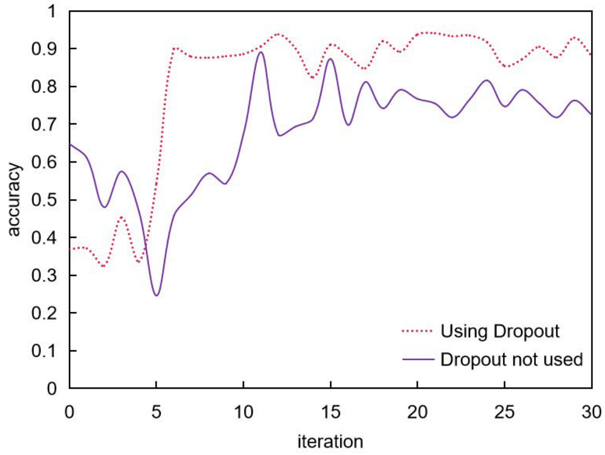

FIGURE 7. Dropout layer contrast experiment.

the human motion video recorded by the camera, in general, the motion changes slowly. Choosing a larger sampling frame interval can help to eliminate redundant information and get more representative motion characteristics. The sampling frame interval is set to 5.

The results of motion classification are obtained by using soft maximum likelihood method. On this basis, an optimization algorithm based on random gradient is proposed in this paper. The batch_size value is set to 32, the initial learning rate is set to 0.002, and the exit rate is set to 0.85. The results of training and testing are shown in Figures 5 and 6.

Visualization results play an important role in various types of emotion recognition, helping to have a clearer understanding of the performance and behavior of the model. Firstly, by visualizing the results, we can observe the distribution of different emotion categories in the feature space and understand whether they have obvious clustering trends. This helps us determine whether the model can effectively distinguish various emotions and whether there are overlapping or confusing situations. Secondly, the visualization results can also demonstrate the model's level of recognition for different emotions. By observing the confusion matrix or the distance between categories, we can determine which emotions are more easily recognized by the model and which may pose challenges. This helps to further improve the model and improve the accuracy of difficult emotions. In addition, visualization can also be used to observe the emotional recognition performance of the model in different contexts or time periods. For example, we can create a time series graph to track the trend of emotional changes, or observe the emotional response of the model in different contexts. This helps to understand the feasibility and stability of emotion recognition in different application scenarios.

In summary, visualization results are crucial for various types of emotion recognition, as they provide intuitive information and insights that help evaluate model performance more comprehensively, improve model design, and provide strong support for emotion analysis in different application fields. When the data set is small, using Dropout can avoid over-fitting to some extent. In this paper, the influence of Dropout on the recognition ability of 3D CNN is tested. In the experiment, the network structure is the 3D CNN model constructed in this paper, and no frame difference channel is added. In the control experiment, two Dropout layers are removed from the 3D CNN model, and the training iteration is 50 times. As shown in fig. 7.

It can be seen that the network identification accuracy with Dropout is obviously higher than that without Dropout, and the convergence speed is faster. The network using Dropout has a certain generalization ability, which can

TABLE 2. Accuracy test results (%).

Number of incorrect actions	Ref[6]	Ref[8]	Proposed method
50	88.13	83.32	94.556
100	70.876	54.697	93.265
150	63.212	48.941	91.58
200	51.438	42.235	91.383
250	49.687	33.84	88.669
300	45.897	32.917	88.459
350	45.799	32.282	88.017
400	30.934	31.041	87.552

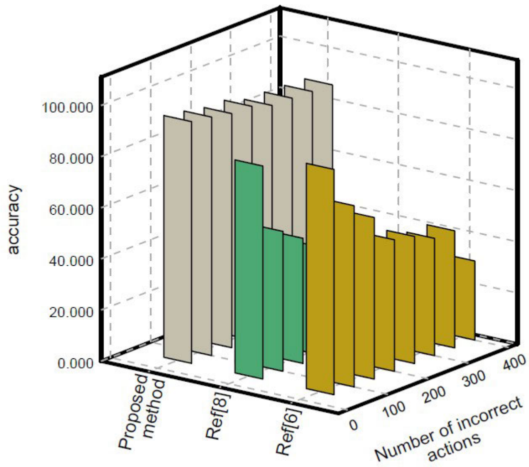

FIGURE 8. Accuracy bar chart.

prevent over-fitting to some extent on the samples with small data volume.

In the experimental study of recognition accuracy, a certain group of image data in the experimental data is randomly selected as the recognition target, different video motion recognition methods are used to recognize the experimental video data, and statistical software is used to calculate and output the recognition accuracy results, as shown in Table 2 and Figure 8.

From the experimental results, it can be clearly observed that as the number of incorrect actions by basketball players increases, the methods [6] and [8] mentioned in the literature show significant shortcomings when dealing with situations that become more complex. Especially when dealing with changes in the direction of basketball players' movements, the recognition effectiveness of these methods has been significantly affected, resulting in a gradual decline in the accuracy of basketball players' incorrect movement recognition. However, compared to this, the method proposed in this article exhibits better robustness in more complex situations. It is worth noting that even if the number of incorrect actions by basketball players reaches 400, the recognition accuracy of this method still maintains a relatively high level of 87.552%. These results indicate that the proposed method can effectively control recognition accuracy in a large range of erroneous actions and complex scenarios, providing a reliable solution for automatic recognition of erroneous actions. This means that whether the number of erroneous actions increases or the situation becomes more complex, the method proposed in this paper can maintain excellent performance and has broad application prospects. These methods have significant industrial significance and can be widely applied in fields such as sports training, health management, and sports medicine. By accurately identifying and correcting athletes' incorrect movements, it can improve sports performance and reduce the risk of sports injuries, which is of great significance for the improvement and protection of athletes. In addition, it can also be used for automated monitoring and evaluation of sports competitions, providing support for the development of the sports industry. In the comparative analysis of performance parameters compared to other datasets, we can see the advantages of our method in processing different datasets compared to the performance of references [6] and [8]. Firstly, as the number of erroneous actions increases, the methods in references [6] and [8] show a significant downward trend in identifying basketball players' erroneous actions. This may be because these methods are unable to effectively capture subtle differences such as changes in motion direction when facing more complex data, resulting in a decrease in recognition accuracy. In contrast, the method proposed in this article maintains high recognition accuracy when processing different datasets. Especially when the number of erroneous actions reaches 400, the accuracy of our method still reaches 87.552%. This indicates that the proposed method has better robustness and can maintain excellent performance in different datasets and more complex scenarios. In addition, the method proposed in this article has wider applicability and can handle different datasets and diverse scenarios. This means that our method can effectively adapt to both the quantity of training data and the diversity of data, and exhibits consistent performance levels across different datasets. The comparative analysis of performance parameters compared to other datasets shows that the method proposed in this paper has better performance and robustness when processing different datasets, providing a reliable solution for automatic recognition of erroneous actions. These results emphasize the practicality and broad application prospects of the method proposed in this paper, especially in dealing with different datasets and complex scenarios, which have important value.

In this paper, some deep learning methods are proposed for action recognition. Although good classification performance has been achieved, there are still many shortcomings. Compared with single action, the interaction between people is often more complicated, and there are more kinds of physical actions involved, such as straight boxing, kicking and blocking in free fighting. Multi-person interaction recognition is still a challenging research topic at present, and it is also a difficult problem to be overcome urgently in the future.

IV. CONCLUSION

In this paper, AI and CV are combined to establish an athlete's false action recognition model based on two channels of 3D

CNN. Simulation results show that the correct rate gradually decreases with the increase of the number of wrong actions. When the number of wrong actions exceeds 400, the correct rate of the proposed method can reach 87.552%. It shows that this method can control the error rate within a reasonable range when identifying athletes' wrong actions. In this paper, some deep learning methods are proposed for action recognition. Although good classification performance has been achieved, there are still many shortcomings. Compared with single action, the interaction between people is often more complicated. Multi-person interaction recognition is still a challenging research topic at present, and it is also a difficult problem to be overcome urgently in the future.

REFERENCES

[1] A. Akula, A. K. Shah, and R. Ghosh, "Deep learning approach for human action recognition in infrared images," *Cognit. Syst. Res.*, vol. 50, pp. 146–154, Aug. 2018.

[2] D. Avola, M. Cascio, L. Cinque, A. Fagioli, and G. L. Foresti, "Affective action and interaction recognition by multi-view representation learning from handcrafted low-level skeleton features," *Int. J. Neural Syst.*, vol. 32, no. 10, Oct. 2022, Art. no. 2250040.

[3] M. Z. Uddin, M. M. Hassan, A. Alsanad, and C. Savaglio, "A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare," *Inf. Fusion*, vol. 55, no. 10, pp. 105–115, 2020.

[4] S. J. Berlin and M. John, "Light weight convolutional models with spiking neural network based human action recognition," *J. Intell. Fuzzy Syst.*, vol. 39, no. 1, pp. 961–973, Jul. 2020.

[5] R. Cui, A. Zhu, G. Hua, H. Yin, and H. Liu, "Multisource learning for skeleton-based action recognition using deep LSTM and CNN," *J. Electron. Imag.*, vol. 27, no. 4, Aug. 2018, Art. no. 043050.

[6] W. Chen, L. Liu, G. Lin, Y. Chen, and J. Wang, "Class structure-aware adversarial loss for cross-domain human action recognition," *IET Image Process.*, vol. 15, no. 14, pp. 3425–3432, Dec. 2021.

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

[8] Y. Hou, L. Wang, R. Sun, Y. Zhang, M. Gu, Y. Zhu, Y. Tong, X. Liu, Z. Wang, J. Xia, Y. Hu, L. Wei, C. Yang, and M. Chen, "Crack-across-pore enabled high-performance flexible pressure sensors for deep neural network enhanced sensing and human action recognition," *ACS Nano*, vol. 16, no. 5, pp. 8358–8369, May 2022.

[9] J. Imran and B. Raman, "Deep motion templates and extreme learning machine for sign language recognition," *Vis. Comput.*, vol. 36, no. 6, pp. 1233–1246, Jun. 2020.

[10] X. Ji, Q. Zhao, J. Cheng, and C. Ma, "Exploiting spatio-temporal representation for 3D human action recognition from depth map sequences," *Knowl.-Based Syst.*, vol. 227, Sep. 2021, Art. no. 107040.

[11] S. Jain, A. Rustagi, S. Saurav, R. Saini, and S. Singh, "Three-dimensional CNN-inspired deep learning architecture for yoga pose recognition in the real-world environment," *Neural Comput. Appl.*, vol. 33, no. 12, pp. 6427–6441, Jun. 2021.

[12] Y. Li, X. Xu, J. Xu, and E. Du, "Bilayer model for cross-view human action recognition based on transfer learning," *J. Electron. Imag.*, vol. 28, no. 3, May 2019, Art. no. 033016.

[13] X. Li, "Human action recognition and athlete's wrong movements detection based on convolutional neural networks," in *Proc. 5th Int. Conf. Control, Autom. Robot. (ICCAR)*, 2019, pp. 387–392.

[14] T. Liu, J. Kong, M. Jiang, and H. Huo, "RGB-D action recognition based on discriminative common structure learning model," *J. Electron. Imag.*, vol. 28, no. 2, p. 15, Mar. 2019.

[15] F. Li, A. Zhu, Z. Liu, Y. Huo, Y. Xu, and G. Hua, "Pyramidal graph convolutional network for skeleton-based human action recognition," *IEEE Sensors J.*, vol. 21, no. 14, pp. 16183–16191, Jul. 2021.

[16] Y. Lin, W. Chi, W. Sun, S. Liu, and D. Fan, "Human action recognition algorithm based on improved ResNet and skeletal keypoints in single image," *Math. Problems Eng.*, vol. 2020, pp. 1–12, Jun. 2020.

[17] M. Li, L. Yan, and Q. Wang, "Group sparse regression-based learning model for real-time depth-based human action prediction," *Math. Problems Eng.*, vol. 2018, pp. 1–7, Dec. 2018.

[18] S. R. Mishra, K. D. Krishna, G. Sanyal, and A. Sarkar, "A feature weighting technique on SVM for human action recognition," *J. Sci. Ind. Res.*, vol. 2020, no. 7, p. 79, 2020.

[19] A. Nuñez-Marcos, G. Azkune, and I. Arganda-Carreras, "Egocentric vision-based action recognition: A survey," *Neurocomputing*, vol. 472, pp. 175–197, Feb. 2022.

[20] H. B. Naeem, F. Murtaza, M. H. Yousaf, and S. A. Velastin, "T-VLAD: Temporal vector of locally aggregated descriptor for multiview human action recognition," *Pattern Recognit. Lett.*, vol. 148, pp. 22–28, Aug. 2021.

[21] J. Ren, N. Reyes, A. Barczak, C. Scogings, and M. Liu, "Toward three-dimensional human action recognition using a convolutional neural network with correctness-vigilant regularizer," *J. Electron. Imag.*, vol. 27, no. 4, Aug. 2018, Art. no. 043040.

[22] L. Shi, "A deep learning-based system for athlete movement error detection in sports training," *IEEE Access*, vol. 8, pp. 189242–189253, 2020.

[23] J. Tang, "An action recognition method for volleyball players using deep learning," *Sci. Program.*, vol. 2021, pp. 1–9, Oct. 2021.

[24] Z. Wang, J. Jin, T. Liu, S. Liu, J. Zhang, S. Chen, Z. Zhang, D. Guo, and Z. Shao, "Understanding human activities in videos: A joint action and interaction learning approach," *Neurocomputing*, vol. 321, pp. 216–226, Dec. 2018.

[25] D. Wang, J. Yang, Y. Zhou, and Z. Zhou, "Human action recognition based on deep network and feature fusion," *Filomat*, vol. 34, no. 15, pp. 4967–4974, 2020.

[26] H. Wang, B. Yu, K. Xia, J. Li, and X. Zuo, "Skeleton edge motion networks for human action recognition," *Neurocomputing*, vol. 423, pp. 1–12, Jan. 2021.

[27] J. Xu and Q. Luo, "Human action recognition based on mixed Gaussian hidden Markov model," in *Proc. MATEC Web Conf.*, 2021, vol. 336, no. 12, p. 06004.

[28] G. Zhang, Y. Rao, C. Wang, W. Zhou, and X. Ji, "A deep learning method for video-based action recognition," *IET Image Process.*, vol. 15, pp. 3498–3511, Dec. 2021.

[29] J. Zhu, W. Zou, Z. Zhu, and Y. Hu, "Convolutional relation network for skeleton-based action recognition," *Neurocomputing*, vol. 370, pp. 109–117, Dec. 2019.

[30] T. Zhang, "Action recognition and athlete's wrong movements detection based on deep learning," in *Proc. 6th Int. Conf. Commun. Inf. Process. (ICCIP)*, 2021, pp. 184–189.