Hindawi
Mobile Information Systems
Volume 2022, Article ID 5168898, 9 pages
https://doi.org/10.1155/2022/5168898

Research Article

A Method of Key Posture Detection and Motion Recognition in Sports Based on Deep Learning

Shaohong Pan

Xichang University, Xichang City, Sichuan 615000, China

Correspondence should be addressed to Shaohong Pan; xcc03300051@xcc.edu.cn

Received 16 February 2022; Revised 20 March 2022; Accepted 21 March 2022; Published 25 April 2022

Academic Editor: Tongguang Ni

Moving target recognition and analysis is an important research direction in the field of computer vision, which is widely used in our life, such as intelligent robot, video surveillance, medical education, sports competition, and national defense security. By analyzing the video of weightlifting, this paper extracts the key postures of athletes' training, so as to assist coaches to train athletes more professionally. Based on DL (Deep Learning), a key pose extraction method of sports video (RoI_KP for short) based on classified learning of regions of interest is proposed. By fine-tuning CNN (Convolutional Neural Network), a network model suitable for video classification of weightlifting in the region of interest is obtained. Finally, according to the classification results, the selection strategy of classification results is designed to extract key poses. According to the characteristics of different modal information, different DNN (deep neural network) is adopted, and various depth networks are combined to mine the multi-modal spatio-temporal depth features of human movements in video. Experimental results show that the method proposed in this paper is very competitive.

1. Introduction

Video analysis technology is the basis of content-based video retrieval. In the traditional text-based query technology, query keywords can basically reflect the query intention [1]. However, in content-based video retrieval, there is a difference between the underlying features and the upper understanding, mainly because the underlying features cannot fully reflect or match the query intention [2]. At the same time, human movements in sports videos are very complicated and skillful, and compared with daily sports [3], the analysis of sports videos is more difficult and challenging [4]. Therefore, the analysis of sports videos can not only bring more viewing effects to sports competitions, but also analyze competitions between athletes and coaches and assist athletes in training. In comparison to manual monitoring, the computer does not need to rest, and it will not miss important information monitoring due to a variety of external factors, allowing it to significantly reduce manpower and material resources while improving work efficiency. This technology is also essential in the medical field. Ward monitoring, for example, can automatically detect abnormal behavior in wards and notify the hospital in real time, which not only reduces the risk of a patient's sudden situation, but also lowers the cost of medical treatment. The query condition in content-based video retrieval is an image sequence or a description of video content. To create an index, first extract the underlying features, then calculate and compare the distance between these features and the query condition to see if they are similar.

Intelligent license plate recognition, real-time 3D effect playback of sports competitions, and other technologies have emerged one after another, which makes the advantages of computer vision task based on machine learning [5–8] more and more obvious, and it has high flexibility and openness, which indicates the development direction of intelligent image processing [9–11]. In the process of weightlifting, whether the athletes' movements are standard or not will directly affect the athletes' performance. In the whole process of weightlifting, there are several key postures that will affect the athletes' performance. These key postures are very important for the success of weightlifting. However, due to the limitation of the hardware conditions of the monitoring equipment (camera) and the influence of

environmental factors, there are many complicated and unavoidable interference factors when acquiring surveillance video, such as various occlusion between objects, noise, background influence, and sudden change of light, which makes it very challenging for us to use target detection and tracking in intelligent monitoring and other related applications. However, the recognition of motion and weight in video is also facing a huge test: the number of super-large-scale videos requires higher and higher computing performance of the algorithm. At the same time, with the emergence of a large number of video content such as aerial photography and first-person perspective, the shortcomings of traditional algorithms in dealing with different camera perspectives, cluttered background, occlusion, and other issues are becoming more and more obvious.

In the training process of athletes, the judgment of key postures in previous weightlifting videos was mainly based on coaches' experience, which not only wasted manpower, but also the extracted key postures were not accurate and easily influenced by subjective factors. Facing weightlifting, this paper applies video analysis technology to weightlifting training, and analyzes the movement of weightlifting process, aiming at putting forward a reliable automatic extraction method of key posture frames in the movement process, which is of far-reaching significance for athletes and coaches to master movement techniques as soon as possible, improve training efficiency, and improve training level.

2. Related Work

Literature [12] detects the edge of the input video stream, uses the edge features to establish gesture descriptors, and then clusters the features calculated by the video stream to get the important gestures. In this way, good model parameters can be learned through loop iterative modeling. Literatures [13, 14] apply SVM (Support Vector Machine) to human motion recognition. They use local spatio-temporal features to represent human motion, and then input local spatio-temporal feature vectors into SVM to judge the category of human motion. Literature [15] extracts spatial features from static points of interest in video, and extracts spatio-temporal features from dynamic points of interest at the same time, and obtains a composite feature set containing both static spatial features and dynamic spatio-temporal features, which is input into linear SVM for classification. A proposed algorithm in the literature [16] is as follows: The gesture features were obtained by training with the edge features obtained after the video was edge detected. The motion was then approved by a majority of the members. Literature [17] proposes a full-body and half-body model, in which the body model employs densely sampled shape context descriptors and the prior model is exhaustively trained in the database using multi-angle and multi-pose. Finally, the image's features are used to detect human body parts in the image, which is determined by the gesture space representation and inference algorithm.

Literature [18] introduced the spatial pyramid pool layer into CNN (Convolutional Neural Network) and proposed SPPnet, which reduced the limitation of CNN network on the size of input pictures and improved the accuracy. On this basis, literature [19] puts forward a faster detection method of Faster-RCNN, and puts forward a new innovation, that is, extracting the RPN (Region Proposal Network) of candidate regions, thus completing the technical evolution from RCNN to Faster-RCNN. Literature [20] puts forward a new detection method, and its performance is greatly improved compared with the traditional one-stage and two-stage framework, especially in the real-time situation, and its accuracy is much higher than YOLOV3 at the same rate. Literatures [21, 22] use NN (nearest neighbor) classifier to judge whether tracking is successful or not. Neural network classifier can measure the similarity between the collected correct image and the current new target image. The DNN (deep neural network) and detector in the tracking-learning-detection framework were improved in literature [23]. The tracking failure point is identified by calculating the DNN's forward-backward error and the output's spatial-temporal similarity. Literature [24] uses the twin network to directly calculate the representation error of the target appearance model, with the goal of selecting the region that is closest to the target model as the tracking result. HRNet is an end-to-end multi-person 2D pose estimation method proposed in the literature [25]. The top-down method has some flaws, such as misalignment of key points in crowded situations or inaccurate detection in bust situations, based on its characteristics. According to the literature [26], organizing two-dimensional posture nodes by guiding all pelvic joints can be done quickly and with good results. If the person in the video is performing fast and complex actions, the estimated two-dimensional posture will be unstable, resulting in motion distortion.

3. Research Method

3.1. Sports Key Posture Detection. In order to detect the target from the image in any scene, the key is to choose a robust feature set, and apply this feature set, so that the targets of different classes have good discrimination and can adapt to the differences of the same class of targets. The features used for target recognition mainly include shape, texture, and color, while the shape of the target is not affected by factors such as illumination, and it is often used as a good feature for target recognition. Different descriptions of the shapes of objects produce different feature sets.

The process of weightlifting is generally divided into five stages: extending the knee to lift the bell, leading the knee to lift the bell, exerting strength, squatting and supporting, and standing up. Coaches and athletes want to observe the movement process of each movement, and hope to observe biomechanical parameters such as the movement track of each snatch or clean and jerk, the speed of each stage, and the performance of work. It is impossible to use traditional pose estimation methods to precisely locate the positions of various parts of the human body, and then extract the motion key frames according to the attitude estimation method. Moreover, there is a serious problem—the similarity between frames. Because of the continuity of motion, the difference between adjacent video frames is very small.

In this chapter, a key pose extraction method of sports video (RoI_KP for short) based on region of interest classification learning is proposed. The algorithm flow is shown in Figure 1.

First, frame the video. Then, based on the first frame of the video, the trained model is used to segment and extract the foreground, and the video is segmented according to the standard of the first frame. The region of interest directly related to key frame extraction is obtained. Furthermore, CNN network is used to extract and classify the features of each video frame to obtain candidate key frames. Finally, the key frame extraction strategy is formulated, and key frames are selected according to the probability value of the corresponding class output by each frame.

Because there is a lot of background interference information in weightlifting video, this paper extracts the region of interest to reduce the influence of background on key frame extraction. Then, divide and fine-tune the region: traverse the image with four corners of the image as the starting points in turn, and update the label value of each pixel point to the minimum value of the label values of four neighboring points that is not zero until the label values of all points do not change;

$$\text{label}_{x,y} = \min\left(\text{label}\middle|\text{label} \in \text{LABEL}_{x,y}\right). \tag{1}$$

Up to this point, a new annotation map of the region of interest has been obtained, and each non-zero and continuous region has a different number. Finally, only the region with the largest number needs to be selected as the region of interest.

Unlike 2D human pose estimation, 3D human pose estimation based on DL is more challenging. This is mainly because 2D pose estimation has a wider training data set, so it can better solve occlusion and accuracy, while 3D pose estimation has great challenges in occlusion and accuracy due to the lack of training data set.

Three-dimensional human posture estimation is to estimate the three-dimensional coordinates (x, y, z) of the related nodes from pictures or videos, which is essentially a regression problem. It is widely used in motion capture system, animation, behavior understanding, and games. It can also be used as an auxiliary link of other algorithms such as pedestrian recognition, and it can also be combined with other tasks related to human body such as human body analysis.

Each block is based on a simple, deep, and multi-layer neural network with batch normalization function. After obtaining 2D key points, input them into 3DPoseNet and output the estimated 3D coordinate key points, each of which is expressed as $P_i = |X_i||Y_i||Z_i|$. Assuming that 3D joint ground truth is available, 3D joint MSE (Mean Square Error) loss is adopted as the loss function by 3D 3DPoseNet:

$$L_{MES} = \sum \left\| P_i - P_i' \right\|_2^2. \tag{2}$$

Among them, $P_i'$is the marked true coordinate of the 3D skeleton, while P_i is the corresponding predicted coordinate, and the coordinates of key points of the 3D skeleton predicted by 3DPoseNet.

In this paper, the loss function with symmetric constraint is used, and the loss function defined in this paper is formula (3).

$$L_{\text{sym}} = \sum_{(i,j)\in E} \left(\left\| P_i - P_j \right\|_2^2 - \left\| P_i' - P_j' \right\|_2^2 \right), \tag{3}$$

where E is the set of all adjacent points, and P_i', P_j' represents the key point of the symmetric part. Of course, there are other restrictions on human bones, such as the limitation of joint angle.

This paper presents a general solution to the occlusion problem that makes extensive use of the torso, limbs, and head. Because 2D posture can accurately predict 3D posture, the restored 3D posture can be divided into four states, including whether the limbs are covered and whether the posture of the entire limbs can be predicted. This paper introduces the developing concept of time series information [17] in this case. The difference is that this paper only needs to determine some position data, i.e., the unpredictable (occluded part) joint position can be determined using the joint points from the previous moment.

$$P_i^t = \lambda P_i^{t-1}. \tag{4}$$

In which, the position t of the joint point i (occlusion node) of the predicted P_i^t at the time point is called $(|X_i||Y_i||Z_i|)^T$, and P_i^{t-1} represents the position at $t-1$ time. λ is a vector, which represents the general goal of the movement direction of people as objects.

3.2. Motion Recognition Method Based on DL. With the rapid development of computer hardware and the successful application of DL in various fields of computer vision, moving target recognition based on DL has also developed into a key technology in the field of computer vision, which is widely used in medical treatment, transportation, security, and so on, and can also be used as the basis of other application technologies, such as image processing and three-dimensional modeling. The traditional methods of target detection can be divided into three steps: selecting candidate areas based on images, extracting visual features, and classifying by a class of commonly used classifiers such as SVM model. With the development and application of DL, it not only simplifies the complexity of traditional methods, but also improves the detection performance. This kind of target detection methods is generally based on region extraction, which can be divided into two types: one-stage method and two-stage method.

Obtaining 3D coordinates from video is extremely difficult. Another method is to represent human motion directly using 2D human posture changes. Poses of the two-dimensional human body in different viewpoints have different forms. Viewpoint normalization, multi-viewpoint traversal, and viewpoint invariant feature extraction are three common techniques for achieving viewpoint-independent representation of human motion. The camera's position

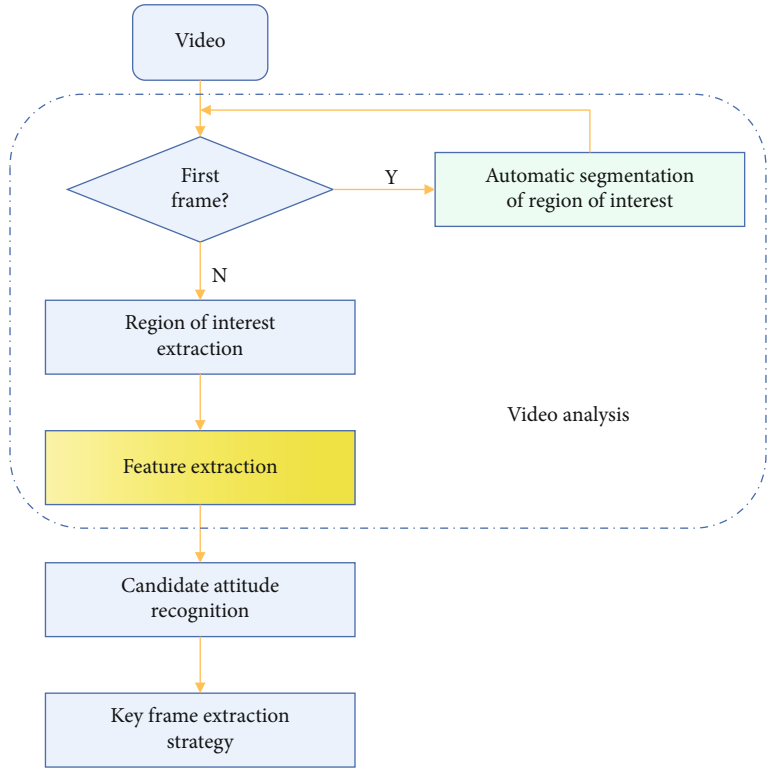

Figure 1: Algorithm framework.

should be fixed, but this is not always the case in sports videos, because the target moves very quickly, and the camera often moves with it to keep up with the target's movement, or the moving target will disappear. When filming a diving video, for example, the camera will move down as the human body falls. This chapter proposes a multi-modal information-based method for action recognition. Figure 2 depicts the general framework of this method.

The static information used in this paper is RGB map and depth map. RGB map contains global information of environment and human body, which can be used to eliminate background clutter and other problems. On the basis of the learning model based on DNN, different network structures are designed for different modes, so as to discover the temporal and spatial depth characteristics of motion video and improve the performance of motion recognition model.

In this chapter, CNN network with the same structure is used for RGB map and depth map. CNN for RGB map represents the color and texture features of human body and background in the video, while CNN for depth map accurately distinguishes the front and back scenes of the video to prevent the misunderstanding of features caused by background interference.

At the top of the network, this chapter uses Softmax loss function. Assuming that the output of the last full connection layer is a vector x_i of K dimension, the Softmax function is defined as:

$$\sigma(x_i) = \exp(x_i) \frac{1}{\sum_{i=1}^{K} \exp(x_i)}. \tag{5}$$

Independent RGB map features and depth map features can be obtained. On the basis of this model, the output of the first fully connected layer can be regarded as a convolution feature with good representation effect, and it has classification and discrimination characteristics.

Human motion in video has time characteristics, so only mining spatial depth features cannot express the temporal characteristics of motion. The networks that deal with timing information in common DL methods are all based on the architecture of RNN (recurrent neural network). In the task of human motion recognition, the trajectory of key skeleton points has effective time information.

RNN network processes sequence data in an iterative way according to the time scale of the sequence. This processing method makes RNN have obvious advantages in modeling and feature extraction of sequence data. This method uses the cross entropy loss function:

$$E\left(y_t, y_t'\right) = -y_t \log y_t' = -\sum_t y_t \log y_t' \tag{6}$$

where y_t is the correct label at time t, and $y_t'$is the network prediction label. The training goal is to make the value of the loss function reach a lower level by calculating the gradient optimization parameter W of the loss function.

The static model and the dynamic model learn the spatial and temporal characteristics of the action video, respectively. Compared with single modal information, the features under multi-modal information have more differences and complementarities, so the fusion of spatial and

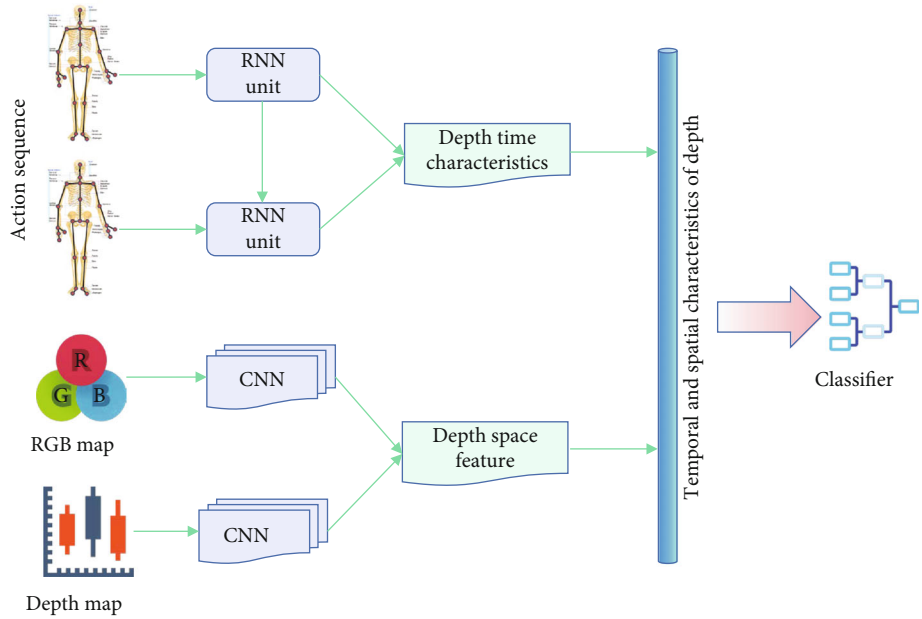

Figure 2: Architecture diagram of deep space-time feature motion recognition.

temporal features can provide information with stronger representation ability. The goal of this experiment is to directly concatenate temporal and spatial depth features and estimate the weights of the two features in a linear combination process based on the test accuracy of the two network models, in order to identify the importance of different features.

In the late stage of fusion, the probability outputs using spatial and temporal depth features are superimposed by linear weighting to get the final predicted value. Specifically, the late fusion adopts the following methods:

$$P = \frac{1}{N}\left(\sum_{i=1}^{N} \alpha P_1 + (1-\alpha)P_2\right) \tag{7}$$

α is the weighting parameter, P_1, P_2 is the output probability of the network using spatial depth feature and temporal depth feature, respectively, and P is the final prediction probability, where N is the number of sample features of the video after multiple sampling.

4. Results Analysis and Discussion

The action behavior of human body is not only distinguishable in image space, but also distinguishable in time series. The tasks of image recognition and detection need to mine the spatial features, while the video increases the information of time dimension relative to the image. Therefore, a video human motion recognition algorithm needs to dig deep into its temporal and spatial features. A segment of video contains many frames.

If all the video tilts are used for a video, it will be a task that consumes a lot of computing time. At the same time, because not all frames are related, the recognition effect will be reduced. Therefore, it is a very important task to find the most distinctive spatio-temporal features in video, which can improve the accuracy of the above algorithms.

Testing with 512 image frames, the correct recognition rate of the four poses is shown in Figure 3.

The experimental results show that the algorithm presented in this paper achieves a good recognition result for each category. The incorrectly divided frame image of a knee joint posture is output along with the probability value for each category. All of the incorrectly divided results are found to belong to the first category of knee joint or the third category of force posture. Different scenes may have different backgrounds or even the same actions. At the same time, it should be noted that if the human body's clothing matches the video's background color, problems such as video exposure caused by outdoor lighting and weather will make it difficult to distinguish actions and behaviors. The same video may not have a static background, and the task will become more difficult as the background changes. Figure 4 depicts the probability change curves of four categories for each video frame.

The RoI_KP method proposed in this paper has greatly improved the performance. It can be seen. Traditional CNN classification algorithm will have great fluctuation in classification performance, and it will be unstable in classification.

With the emergence of various new shooting methods, such as self-timer, aerial photography, and first-angle shooting, it is impossible for video information to always keep the same camera angle. Therefore, the same video action sequence may cause different feature representations due to different camera angles. Therefore, the identification of different camera angles is also an urgent problem to be solved. The experimental input is set as a continuous and orderly video meal, so the representation of features has all the

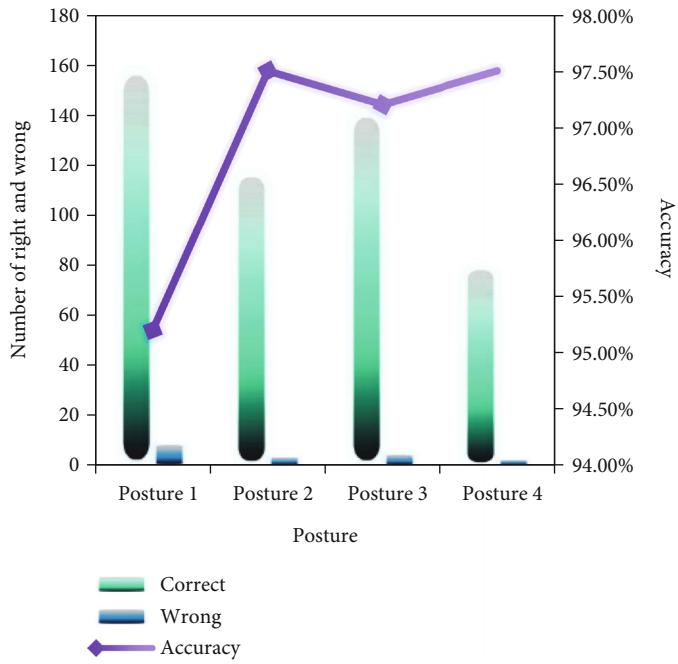

Figure 3: Test accuracy of four key frames.

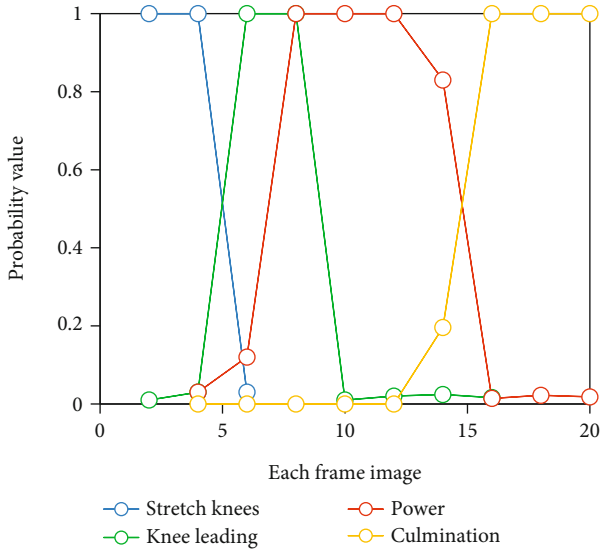

Figure 4: Test video results.

spatial information, and at the same time contains part of the temporal information. However, there is little pre-training information of depth map information, so the network is trained from scratch. Figures 5 and 6 are the comparison results of two groups of videos.

While the traditional CNN algorithm shows great fluctuation in the fourth key frame and the probability values of other poses are relatively large, which has an impact on the classification results, the traditional CNN algorithm has a good effect in the fourth key frame and the probability value of the fourth key pose is very large. The pre-training process is not used for RNN but used for bone point information because the processed features obtained after pre-processing are sufficient for training. Learning transfer allows researchers to obtain good training results even when the new data set is small, labeling is incomplete or small, and the distribution of the training and test sets of data is different. Low-level features can also be used for similar tasks because the representation of convolution kernels for CNN is a process from low-level to high-level and from general to special.

In this paper, the 3DPoseNet network is evaluated according to the public data set Human3.6M, and the other two data sets are used for testing. Human3.6M is the most widely used public data set in 3D pose at present, which contains 3.6 million video frames for 11 topics, among which 7 topics use 3D annotation pose, each object performs 15 actions, and the video frames (video with frame rate of 50fps) are captured by four cameras with different angles of view, as shown in Figure 7.

It demonstrates that the 3DPoseNet model's results have a lower average error than all other methods and are unaffected by other data or operation methods. The average error of MPJPE in the 3DPoseNet network of this method is reduced by at least 6 mm, demonstrating the method's effectiveness. In general, if a human posture is similar at two different moments of an action, the distance between the two descriptor features describing the human posture will be small; on the other hand, if the human posture is very different at two different moments, the distance between the two descriptor features will be large, and this property has little relation with the viewpoint position of observing the movement. The self-similar matrix with the same characteristics has similar patterns, as shown in Figure 7, indicating that changes in sports performers have little impact on the self-similar matrix. The difference in self-similar matrix based on different features is very obvious from a vertical

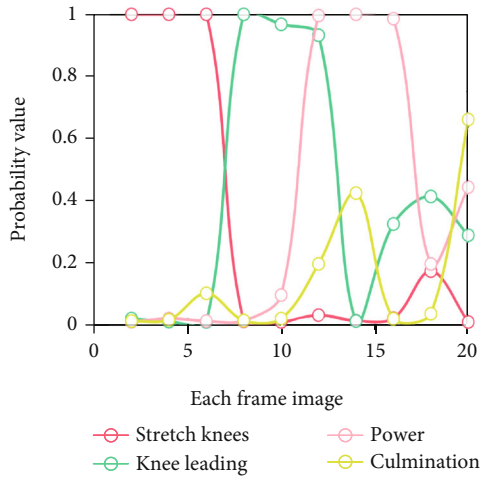

Figure 5: CNN test results.

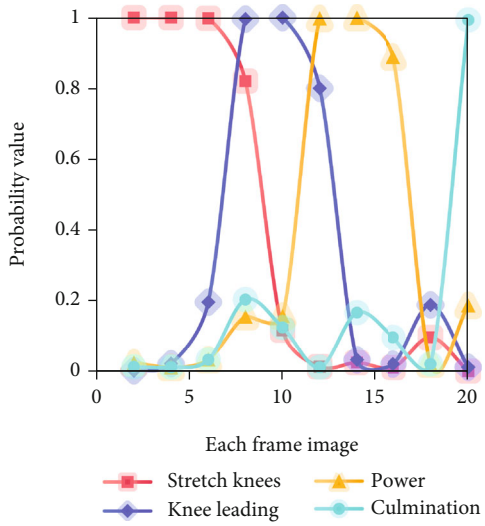

Figure 6: Results based on RoI_KP.

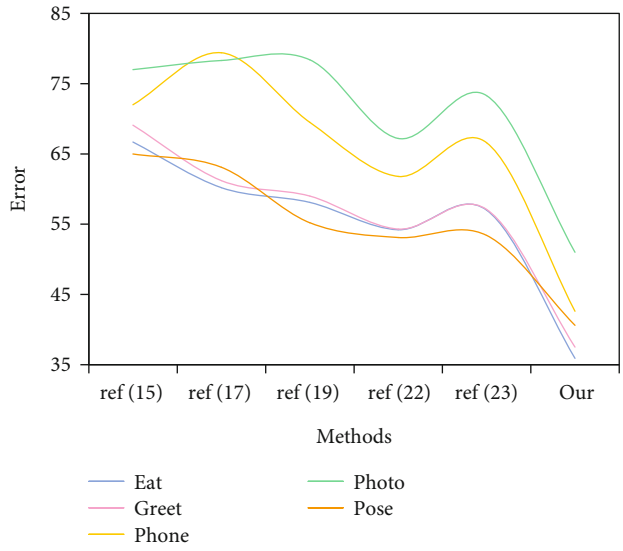

Figure 7: Results of average position error per joint.

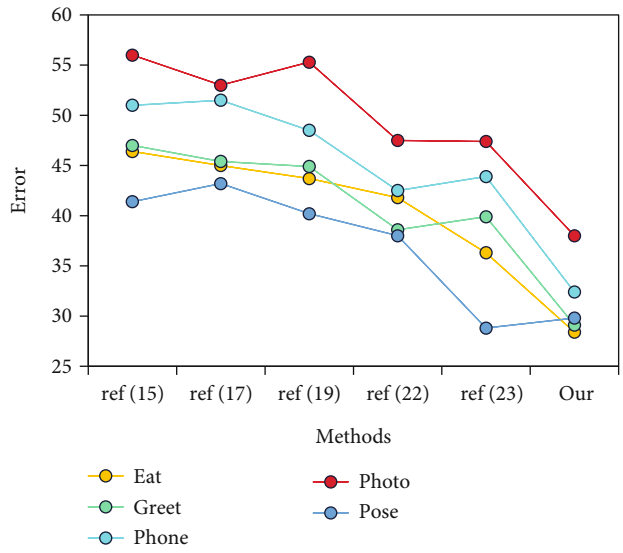

Figure 8: Joint position error based on Procrustes analysis.

perspective, so the self-similar matrix is dependent on the image's underlying features. As shown in Figure 8, it can be clearly found that the accuracy of this method is greatly improved.

All the videos of human body movements were shot in fixed scenes. In order to prevent the moving target from losing from the field of vision, moving lens is often used to shoot, and it is difficult to detect the movement of human body from the optical flow field. Figure 9 shows the process of 3D CNN in pre-training stage and network fine-tuning stage, respectively. It can be seen that the fluctuation and loss output are larger in the training process, and the network output loss is reduced after fine-tuning, thus improving the final classification effect of the model and reducing the time cost of network learning.

The action recognition method based on DL in single mode includes CNN and recurrent neural network. It also explains the problems and difficulties brought by single-mode video motion recognition, and introduces multi-mode motion recognition method. DL gets better results based on a large amount of training data. Different depth networks have different characteristics. CNN network pays more attention to the relationship between local information, so it is suitable for image recognition and detection tasks.

Experiments show that the self-similar matrix of human posture obtained by using directional gradient histogram of human foreground image to represent human posture in different viewpoints has better stability. By applying the filter to the post-processing of 3D pose attitude, the predicted attitude stability can be effectively improved. On the basis of successful motion capture, through the analysis and extraction of human motion feature parameters, we can automatically analyze and understand various human motions and behaviors. Motion recognition technology has a wide application prospect and great economic and social value in

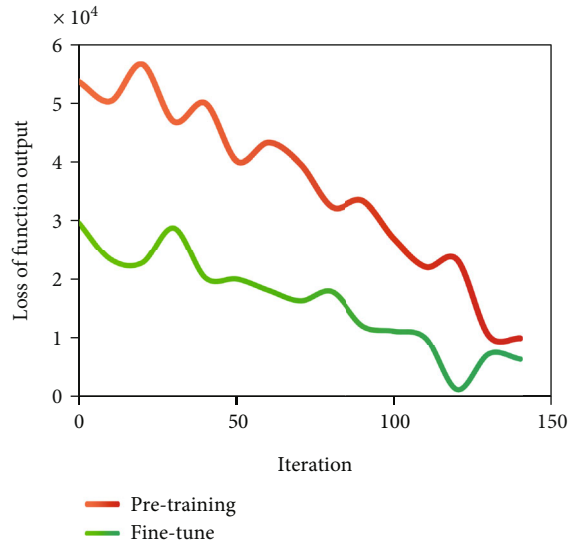

Figure 9: 3-D CNN pre-training and fine-tuning network loss.

advanced human-computer interaction, rehabilitation engineering, sports analysis, somatosensory game control, and content-based retrieval.

5. Conclusion

As the most important element in the human environment, rich and diverse human movements carry a great deal of information which is extremely important for human social interaction. Therefore, studying human movements has profound economic and social value. The research of human motion recognition based on vision involves multidisciplinary knowledge, and integrates the related research results of computer vision, pattern recognition, image processing, machine learning, and many other disciplines. When we regress 2D to 3D pose nodes based on DL, we can effectively improve the accuracy of the network model by calculating the scale of 2D to 3D node coordinates of each frame based on the least square method. The human body in weightlifting video is extracted by skeleton to enhance the expression of features, thus further improving the accuracy of key frame extraction. Deep CNN is used to mine the spatial characteristics of static information, improve the representation of dynamic information, and recursive neural network is used to process it. Experimental results show that the method proposed in this paper is very competitive.

From 2D pose estimation to 3D pose prediction, if there is occlusion, it is mainly self-occlusion, which will greatly affect people's visual effect. Therefore, the focus of our future work is to better solve the occlusion problem and improve the processing speed of the framework.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author does not have any possible conflicts of interest.

References

[1] J. Miller, U. Nair, R. Ramachandran, and M. Maskey, "Detection of transverse cirrus bands in satellite imagery using deep learning," *Computers & Geosciences*, vol. 118, pp. 79–85, 2018.

[2] H. Liu, J. Jin, Z. Xu, Y. Bu, Y. Zou, and L. Zhang, "Deep learning based code smell detection," *IEEE Transactions on Software Engineering*, vol. 99, 2021.

[3] M. Gao, W. Cai, and R. Liu, "AGTH-Net: attention-based graph convolution-guided third-order hourglass network for sports video classification," *Engineering*, vol. 2021, pp. 1–10, 2021.

[4] Z. Wang, Z. Zhou, H. Lu, and J. Jiang, "Global and local sensitivity guided key salient object re-augmentation for video saliency detection," *Pattern Recognition*, vol. 103, no. 2, article 107275, 2020.

[5] J. Zhou, Y. Wang, and W. Zhang, "Underwater image restoration via information distribution and light scattering prior," *Computers and Electrical Engineering*, vol. 100, article 107908, 2022.

[6] X. Gu, W. Cai, M. Gao, Y. Jiang, X. Ning, and P. Qian, "Multi-Source Domain Transfer Discriminative Dictionary Learning Modeling for Electroencephalogram-Based Emotion Recognition," in *IEEE Transactions on Computational Social Systems*, pp. 1–9, Institute of Electrical and Electronics Engineers, 2022.

[7] M. Zhao, A. Jha, Q. Liu et al., "Faster mean-shift: GPU-accelerated clustering for cosine embedding-based cell segmentation and tracking," *Medical Image Analysis*, vol. 71, article 102048, 2021.

[8] W. Cai, B. Zhai, Y. Liu, R. Liu, and X. Ning, "Quadratic polynomial guided fuzzy C-means and dual attention mechanism for medical image segmentation," *Displays*, vol. 70, article 102106, 2021.

[9] J. Zhou, X. Wei, J. Shi, W. Chu, and W. Zhang, "Underwater image enhancement method with light scattering characteristics," *Computers and Electrical Engineering*, vol. 100, article 107898, 2022.

[10] D. Yao, Z. Zhi-li, Z. Xiao-feng et al., "Deep hybrid: multi-graph neural network collaboration for hyperspectral image classification," *Defence Technology*, vol. 5, 2022.

[11] J. Zhou, D. Liu, X. Xie, and W. Zhang, "Underwater image restoration by red channel compensation and underwater median dark channel prior," *Applied Optics*, vol. 61, no. 10, pp. 2915–2922, 2022.

[12] A. Lamas, S. Tabik, P. Cruz et al., "MonuMAI: dataset, deep learning pipeline and citizen science based app for monumental heritage taxonomy and classification," *Neurocomputing*, vol. 420, pp. 266–280, 2021.

[13] S. Nandyal and S. L. Kattimani, "Bird swarm optimization-based stacked autoencoder deep learning for umpire detection and classification," *Scalable Computing*, vol. 21, no. 2, pp. 173–188, 2020.

[14] X. Zhao, T. Zuo, and X. Hu, "OFM-SLAM: a visual semantic SLAM for dynamic indoor environments," *Mathematical Problems in Engineering*, vol. 2021, 16 pages, 2021.

[15] Q. Guo, "Detection of head raising rate of students in classroom based on head posture recognition," *Traitement du Signal*, vol. 37, no. 5, pp. 823–830, 2020.

[16] P. Wang, "Research on sports training action recognition based on deep learning," *Scientific Programming*, vol. 2021, Article ID 6878, 8 pages, 2021.

[17] A. Bruno, F. Gugliuzza, R. Pirrone, and E. Ardizzone, "A multi-scale colour and keypoint density-based approach for visual saliency detection," *Access*, vol. 8, pp. 121330–121343, 2020.

[18] Y. Xu, "A sports training video classification model based on deep learning," *Scientific Programming*, vol. 2021, Article ID 2896, 11 pages, 2021.

[19] J. Liu, D. Chen, Y. Wu, R. Chen, P. Yang, and H. Zhang, "Image edge recognition of virtual reality scene based on multi-operator dynamic weight detection," *Access*, vol. 8, pp. 111289–111302, 2020.

[20] F. A. Memon, U. A. Khan, A. Shaikh, A. Alghamdi, P. Kumar, and M. Alrizq, "Predicting actions in videos and action-based segmentation using deep learning," *Access*, vol. 9, pp. 106918–106932, 2021.

[21] G. Kalakoti and G. Prabakaran, "Key-frame detection and video retrieval based on DC coefficient-based cosine orthogonality and multivariate statistical tests," *Traitement du Signal*, vol. 37, no. 5, pp. 773–784, 2020.

[22] H. Gu, J. Zhang, T. Liu et al., "DIAVA: a traffic-based framework for detection of SQL injection attacks and vulnerability analysis of leaked data," *IEEE Transactions on Reliability*, vol. 69, no. 1, pp. 188–202, 2020.

[23] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[24] B. Pu, N. Zhu, K. Li, and S. Li, "Fetal cardiac cycle detection in multi-resource echocardiograms using hybrid classification framework," *Future Generation Computer Systems*, vol. 115, no. 3, pp. 825–836, 2021.

[25] J. Cui, M. Wang, Y. Luo, and H. Zhong, "DDoS detection and defense mechanism based on cognitive-inspired computing in SDN," *Future Generation Computer Systems*, vol. 97, pp. 275–283, 2019.

[26] C. Li-quan, L. You, F. Shen, Z. Shan, and J. Chen, "Pose recognition in sports scenes based on deep learning skeleton sequence model," *Journal of Intelligent and Fuzzy Systems*, vol. 3, pp. 1–10, 2021.