Research on functional architecture of sports application system based on AI technology

Zhiqiang Min*
Nanchang Institute of Technology, Nanchang 330044, Jiangxi, China.

ABSTRACT

In the field of sports analysis, combining sports action recognition with sports events and analyzing the athletes' sports situation in the video can provide athletes with auxiliary training, find their strengths and weaknesses, optimize the training system and promote their growth. Aiming at the low accuracy of sports video action recognition method, this paper applies the deep learning (DL) technology in artificial intelligence (AI) to sports action recognition, constructs the functional architecture of sports application system, and conducts comparative experiments on standard sports video action data sets to verify the feasibility and effectiveness of the proposed algorithm from the recognition accuracy and response time respectively. The test results show that the proposed method is superior to the traditional motion recognition model in accuracy and efficiency.

Keywords: Sports; Motion recognition; Artificial intelligence; Deep learning

1. INTRODUCTION

Motion recognition of sports videos is a multi-category pattern recognition problem, which mainly faces two challenges: one is to extract effective features from similar sports actions in various sports videos; The other is to construct a machine learning model to complete the classification of action features [1]. In the stage of physical training, some postures are not in place, which leads to the decline of training quality and ultimately affects students' physical performance [2]. The general sports action recognition method includes two steps: feature extraction and classifier design. Firstly, the feature of the original input data is extracted by a manually designed feature extractor, and then the extracted features are trained on the classifier to classify and recognize the actions [3-4]. In the field of sports analysis, combining sports action recognition with sports events and analyzing the athletes' sports situation in the video can provide athletes with auxiliary training, find their strengths and weaknesses, optimize the training system, promote their growth, and finally make athletes achieve better training results [5].

In order to improve the effect of physical training, it is necessary to identify the human posture in the training process in order to correct the problems in physical training in time. Traditional action recognition methods are generally suitable for small-scale data sets with few types of actions, small amount of data and low complexity of actions [6]. However, with the advent of the era of big data, multimedia information data is growing exponentially, and the traditional method based on artificial design features can not adapt to the current large-scale data. Since the rise of DL method, DL method has been introduced to automatically extract motion features, reducing the cost of manual feature design [7]. A large quantity of motion data make the recognition model obtained by training more accurate, and the development of computing power makes the model training no longer stretched. Ou et al. extracted the motion trajectory of joint points through three-dimensional modeling [8]. Rodrigues et al. used optical flow method to extract gait cycle, obtained gait cycle energy map and direction gradient histogram, and recognized gait through feature fusion [9]. Dong et al. used Gaussian mixture model and support vector machine to complete sports action recognition. However, the feature dimension of Gaussian mixture model used in this method is too high, which is not conducive to the classification of support vector machine, so the effect of sports action recognition is poor [10]. In this paper, DL technology in AI is applied to sports action recognition, and the functional architecture of sports application system is constructed.

*E-mail: 56857798@qq.com

Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023), edited by Srikanta Patnaik,
Tao Shen, Proc. of SPIE Vol. 12779, 127791H · © 2023 SPIE · 0277-786X · Published under a
Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2689454

2. METHODOLOGY

2.1 Image sampling and feature expression of human posture in physical training

When recognizing the actions of multiple people in a video sequence, the human motion images taken by multi-view cameras are blocked and crossed, so it is necessary to identify the human motion trajectory first. The rapid development of DL has injected new impetus into the research of motion recognition. DL is generally built into a multi-level structure composed of nonlinear computing units, and the output of the low-level network is used as the input of the high-level network. This kind of structure can often extract target features from a large quantity of data, but ordinary machine learning needs to define the feature quantity artificially, and then train the defined feature quantity [11]. In DL-based motion recognition methods, most of them use end-to-end methods to extract motion features and classify them, in which depth CNN plays an important role. Appropriate CNN can capture abundant action discrimination features and make the model have better learning and representation ability. Over-fitting often occurs in this kind of complex networks. Optimizing the network structure and improving the training algorithm to reduce over-fitting and improve the generalization performance are also problems to be solved. A single depth structure has a good performance for a specific problem, and combining multiple depth structures to solve multi-characteristic problems has become the current research trend of DL. The functional architecture of sports application system is shown in Figure 1.

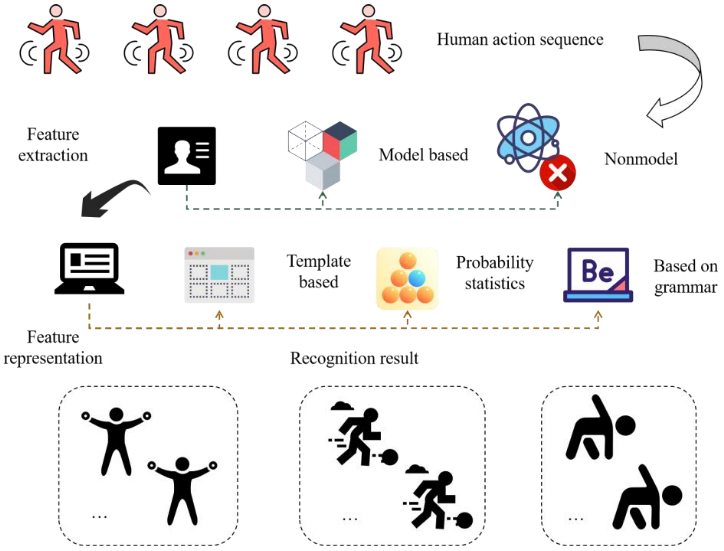

Figure 1. Functional architecture of sports application system

When the quantity of samples in the data set is limited, it is a common practice to choose an appropriate data enhancement strategy to increase the quantity of samples and improve the diversity of samples, thus alleviating the over-fitting problem and improving the robustness of the model. DL comes from machine learning, and its essence is also a method to find the optimal solution. Data is passed through some nonlinear structures and these nonlinear structures are changed by generation. After many iterations, this nonlinear structure becomes the optimization function of this kind of data. However, the driving mode of DL is different from that of traditional machine learning, which is strictly proved by statistics and mathematics. This method is driven by a large quantity of data. Through different network structures, the characteristics that can be used to express the data pointing results are found in the data according to the set rules, and the model parameters are optimized according to these characteristics.

2.2 Sports action recognition algorithm

The basis of sports action recognition is to obtain more discriminating time and space information, and deeper CNN can extract more discriminating information, thus improving the prediction performance [12]. In general, the method of improving the accuracy of the model will deepen or broaden the scale and parameters of the network, but this will not only increase the quantity of super-parameters, but also increase the difficulty of network design and computational overhead, and also make the network prone to over-fitting. Generally speaking, people can show a variety of motion characteristics when doing actions, but it is not comprehensive to describe human actions with only one feature. Although the types of actions are the same, the trajectory of a single feature of limb joints may be the same or close, for example, a joint point in a frame of different actions will have the same angle. In this case, the joint angle characteristics can not completely distinguish the movements, which may cause classification errors and reduce the recognition rate[13].

For general CNN, its input is single frame or continuous stacked frame, and the limitation of input form makes the network only pay attention to the action changes in a short period of time, and can not effectively extract the time information in a long period of time. Data enhancement refers to making limited data produce value equivalent to more data through image translation, flipping, cropping and other transformation methods without substantial increase of data. Using data enhancement method for action data set can not only increase the diversity of samples, but also reduce the over-fitting phenomenon of the network. In the training process, the network with stronger generalization ability can be obtained by various transformations of pictures, which makes the model more robust. The sports action recognition model based on the improved CNN algorithm is shown in Figure 2.

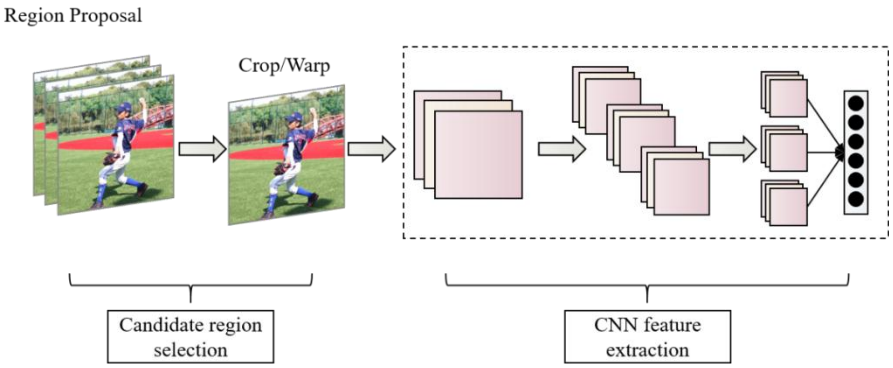

Figure 2. Sports action recognition model

In the application of video-based sports action recognition, there is a strong spatial correlation and time correlation between the video frames of the action. A good action video representation method can not only effectively avoid the limitation of sample number, but also make full use of the spatial and time information in the video with as little redundancy as possible. Suppose the formula of the fully connected layer feature output $x^{(l)}$ of the l layer is as follows:

$$x^{(l)} = f\left(w^{(l)}x^{(l-1)} + b^{(l)}\right) \tag{1}$$

Among them, $w^{(l)}$ represents the weight parameter, and $b^{(l)}$ is the bias term. The Softmax regression classifier needs to iteratively update and learn, and the functions to be learned are:

$$h_w\left(\vec{x}\right) = \frac{1}{\sum_{i=1}^{k} e^{\vec{w_i}\cdot\vec{x}+b_i}} \begin{bmatrix} \vec{w_1}\cdot\vec{x}+b_1 \\ \vec{w_2}\cdot\vec{x}+b_2 \\ \cdots \\ \vec{w_k}\cdot\vec{x}+b_k \end{bmatrix} \tag{2}$$

Among them, k represents the quantity of categories to be classified, and b_i and $\vec{w_i}$ represent the offset vector and weight vector corresponding to the i th category. Equation (3) represents the probability value that the sample $\vec{x}$ is the j class.

$$P\left(y=j\middle|\vec{x}\right)=\frac{\vec{w}_j\cdot\vec{x}+b_j}{\sum_{i=1}^{k}e^{\vec{w}_i\cdot\vec{x}+b_i}}\qquad\sum_{j=1}^{k}P\left(y=j\middle|\vec{x}\right)=1\tag{3}$$

After training and learning to get $\vec{w}_i$ and b_i, the objective loss function can be expressed as:

$$J(w,b)=-\frac{1}{m}\sum_{j=1}^{m}\sum_{l=1}^{k}1\left\{y^{(j)}=l\right\}\log\frac{e^{\vec{w}_l\cdot\vec{x}+b_l}}{\sum_{i=1}^{k}e^{\vec{w}_i\cdot\vec{x}+b_i}}\tag{4}$$

Among them, m represents the quantity of samples in the training set, k represents the quantity of classification categories, and $1\{\cdot\}$ is an indicative function. When $\mathrm{y}^{(\mathrm{j})}=l$, the function value is 1, otherwise it is 0. Optimized by an autoencoder for high-dimensional input data $x\in R^N$:

$$\min_{f,g}\left\|x-f(g(x))\right\|\tag{5}$$

Among them, the encoder $\mathrm{y}=\mathrm{g}(\mathrm{x})$ maps the input data to the low-dimensional space $y\in R^M, N>M$. Let $x\in R^N$ be the observations of time t, the optimization function of the time encoder is as follows:

$$\min_{f,g}\left\|X_{(t+1):(t+\Delta t)}-f\left(g\left(X_{(t-\Delta t+1):t}\right)\right)\right\|\tag{6}$$

Where the encoder $y=g\left(X_{(t-\Delta t+1):t}\right)$ maps the input data to the low-dimensional space $y\in R^M,(N\times\Delta t)>M$. The decoder $\hat{X}_{(t+1):(t+\Delta t)}=f(y)\in R^{N\times\Delta t}$ is used to map back to the data space.

In order to extract richer temporal and spatial features, RGB and optical flow data are used to make the model make full use of the action space and time sequence information in the video. Secondly, in view of the phenomenon that similar actions are easily misjudged in long-lasting videos, the optimal video segmentation value is obtained through testing, so that the model can better distinguish similar actions of molecular action sharing phenomenon, thus better solving some misjudgments caused by similar sub-actions. When using DL method to classify and identify human actions, a large amount of data is needed. But not in all cases, there is enough data to carry out the experiment. Therefore, if you want to have enough data to complete the task, there are generally two methods to choose from. One is to find more marked data resources, and the other is to make full use of existing data for data enhancement.

3. RESULT ANALYSIS AND DISCUSSION

Video data, as the most easily available data format, is also the most commonly used modal data in sports action recognition tasks. Video modal data contains more detailed human movement information, and depicts the movement of human parts more carefully. In addition, for some learning tasks that need to be assisted by scenes or objects, video modal data provides more auxiliary information besides the movement itself. However, some useless interference information is inevitably mixed in the obtained video data, which makes the target information scattered. The combination of the data of the two modes can complement each other and obtain higher accuracy of action recognition. After the motion features pass through the classifier, the feature vectors are mapped into classification score vectors. Taking the index position corresponding to the maximum probability value is the classification result of the model. The accuracy of motion recognition of different algorithms is shown in Figure 3.

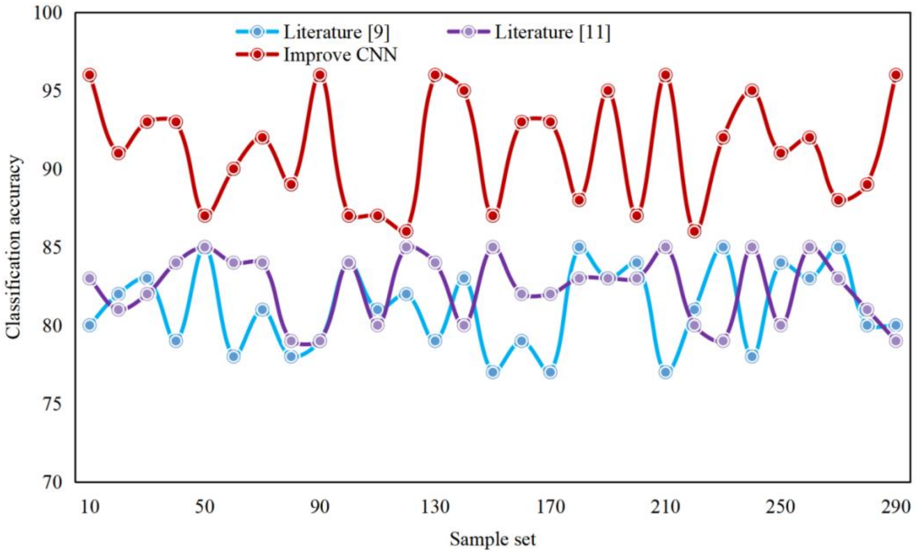

Figure 3. Recognition accuracy of different algorithms

From the angle of information form, compared with RGB image information, the information of bone points is extremely dense, and the information is generally lighter and less redundant, but the overall information is not as high as that of RGB image information. Moreover, when describing the interaction between people and some scenes and objects, the problem of incomplete expression of overall action information will occur due to the lack of description of interactive objects. In the case of a large quantity of unknown input data features, it is not appropriate to specify a unified static structure for different types of action sequence data with different lengths. A lot of attempts must be made to obtain a suitable network structure in order to obtain satisfactory analysis results.

In order to improve the anti-noise performance of the original network while maintaining the performance of feature extraction and classification, the residual shrinkage subnet is added to the input part of the original network. In the face of complex problems, it is often necessary to deepen the network level to fit more complex functions and models to solve such problems. The network level can not be increased indefinitely all the time. Increasing the network level will not only bring redundancy of training parameters, but also lead to slow training and increase the degree of over-fitting. More importantly, the superposition of these levels represents that features are constantly being extracted. However, with the increase of feature depth, some features become no longer obvious, and even important feature information is lost in the network extracted layer by layer, which leads to the degradation of the network.

The core processing stage of CNN is the convolution operation of feature map, which is similar to a filtering form in image processing algorithm. Conventional filter parameters are unchanged, while CNN sets adaptive filter parameters to extract regional features in this way. This feature extraction method has been proved to be spatial sensitive in a large quantity of experiments, which is manifested in the acquisition of neighborhood correlation of pixels in feature maps in image processing. The input video frame sequence will first extract the feature information of video action through the feature extraction module of the network. Then, the feature information extracted from the model is mapped to the marking space of the action sample by using the full connection layer through linear transformation. Finally, the Softmax classifier is used to evaluate the probability of video action categories, and the action category with the highest probability is taken as the recognition result of video action. Figure 4 shows the MAE comparison of different algorithms on the test set. Figure 5 shows the response time comparison of different algorithms.

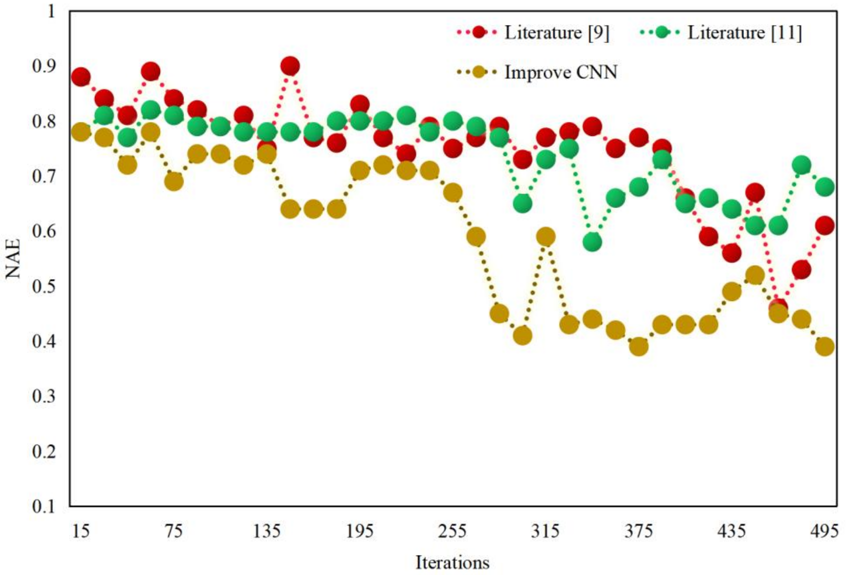

Figure 4. MAE comparison of different algorithms

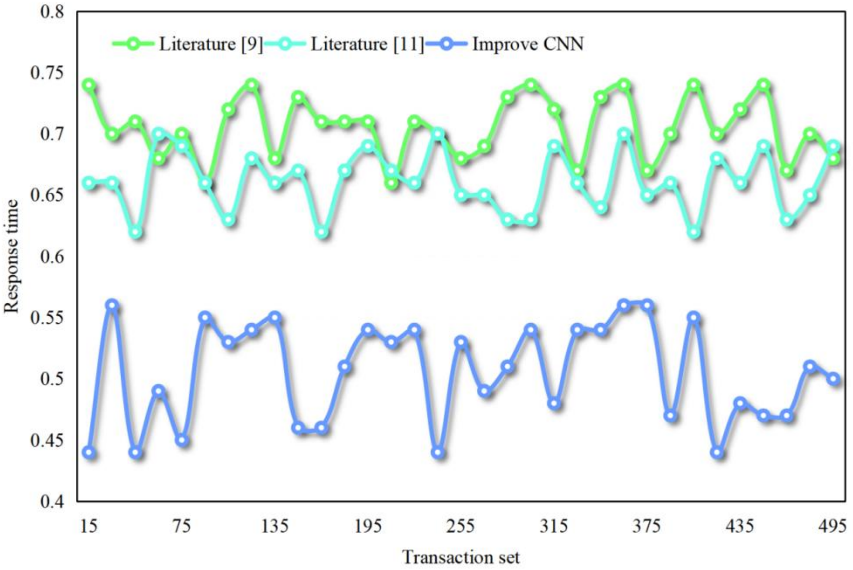

Figure 5. Comparison of response time of different algorithms

When dealing with the closely related problems in time domain, it is generally noticed that the information at different time points has certain correlation, and the information distributed at different time points plays a different proportion in the whole information set. In order to make this correlation more explicit and allocate the proportion of data at different time points in the whole data set more reasonably, the common method is to use attention mechanism. As can be seen from Figure 4 and Figure 5, the improved MAE of CNN on the test set is obviously improved compared with the traditional algorithm, and the error sum and response time are obviously reduced. Compared with high-level statistical features, using the bottom features to build a sports feature recognition model can get higher accuracy. Through the improvement of this paper, the convergence speed of CNN parameters is faster, and the final model classification accuracy is higher.

4. CONCLUSIONS

For multi-modal motion recognition, the main obstacle is how to effectively and correctly model and use the information of each modal motion. In this paper, DL technology in AI is applied to sports action recognition, and the functional architecture of sports application system is constructed. Compared with the traditional algorithm, the improved MAE of CNN on the test set is obviously improved, and the error sum and response time are obviously reduced. In order to obtain more discriminant features for action recognition, the depth and RGB modal data are used for action recognition at the same time, and an adaptive weight multi-stream fusion method is given to realize more effective multi-stream data fusion and improve the action recognition rate. Compared with high-level statistical features, using the bottom features to build a sports feature recognition model can get higher accuracy. If you want to really understand human movements, you need to understand the evolution of the relationship between movements, describe the movements and their constituent elements at the conceptual level, and correctly understand the semantics of movements. Therefore, action semantics can be studied, so that sports action recognition has further development and progress.

ACKNOWLEDGEMENTS

Project Name: Science and Technology Research Project of Jiangxi Provincial Department of Education (Fund number:GJJ212106).

REFERENCES

[1] Yi, Y., Cheng, Y., Xu, C., Mining human movement evolution for complex action recognition. Expert Systems with Applications, vol. 78, no. 7, pp. 259-272 (2017).

[2] Peng, C., Huang, H., Tsoi, A. C., et al., Motion boundary emphasised optical flow method for human action recognition. IET Computer Vision, vol. 2020, no. 6, pp. 14 (2020).

[3] Xu, W., Miao, Z., Yu, J., et al., Action Recognition and Localization with Spatial and Temporal Contexts. Neurocomputing, vol. 333, no. 3, pp. 351-363 (2019).

[4] Lei, Q., Du, J. X., Zhang, H. B., et al., A Survey of Vision-Based Human Action Evaluation Methods. Sensors, vol. 19, no. 19, pp. 4129 (2019).

[5] Liu, J., Che, Y., Action recognition for sports video analysis using part-attention spatio-temporal graph convolutional network. Journal of Electronic Imaging, vol. 30, no. 3, pp. 33017 (2021).

[6] Wu, D., Online position recognition and correction method for sports athletes. Cognitive Systems Research, vol. 52, no. 12, pp. 174-181 (2018).

[7] Nazir, S., Yousaf, M. H., Nebel, J. C., et al., A Bag of Expression framework for improved human action recognition. Pattern recognition letters, vol. 103, no. 2, pp. 39-45 (2018).

[8] Ou, H., Sun, J., Spatiotemporal information deep fusion network with frame attention mechanism for video action recognition. Journal of Electronic Imaging, vol. 28, no. 2, pp. 1 (2019).

[9] Rodrigues, A., Pereira, A. S., Rui, M., et al., Using Artificial Intelligence for Pattern Recognition in a Sports Context. Sensors, vol. 20, no. 11, pp. 3040 (2020).

[10] Dong, J., Yang, W., Yao, Y., et al., Knowledge memorization and generation for action recognition in still images. Pattern Recognition, vol. 120, no. 10, pp. 108188 (2021).

[11] Xu, X., Hospedales, T., Gong, S., Transductive Zero-Shot Action Recognition by Word-Vector Embedding. International Journal of Computer Vision, vol. 123, no. 3, pp. 309-333 (2017).

[12] Lemieux, N., Noumeir, R., A Hierarchical Learning Approach for Human Action Recognition. Sensors, vol. 20, no. 17, pp. 4946 (2020).

[13] Lin, B., Fang, B., Yang, W., et al., Human Action Recognition Based on Spatio-temporal Three-Dimensional Scattering Transform Descriptor and An Improved VLAD Feature Encoding Algorithm. Neurocomputing, vol. 348, no. 7, pp. 145-157 (2018).