IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 34, No. 5, October, 2020, pp. 663-671
Journal homepage: http://iieta.org/journals/ria

Recognition of Wrong Sports Movements Based on Deep Neural Network

Haiying Wang

College of Physical Education, Baoji University of Arts and Sciences, Baoji 721013, China

Corresponding Author Email: why9130@163.com

https://doi.org/10.18280/ria.340518

Received: 21 May 2020
Accepted: 19 September 2020

Keywords:
three-dimensional (3D) convolutional neural network (CNN), demonstrative sports movements, movement standardization, wrong movement recognition

ABSTRACT

During physical education (PE), the teaching quality is severely affected by problems like nonstandard technical movements or wrong demonstrative movements. High-speed photography can capture instantaneous movements that cannot be recognized with naked eyes. Therefore, this technology has been widely used to judge the sprint movements in track and field competitions, and assess the quality of artistic gymnastics. Inspired by three-dimensional (3D) image analysis, this paper proposes a method to recognize the standard and wrong demonstrative sports movements, based on 3D convolutional neural network (CNN) and graph theory. Firstly, a 3D posture perception strategy for demonstrative sports movements was constructed based on video sequence. Next, the authors provided the framework of the recognition system for standard and wrong demonstrative sports movements. After that, a 3D CNN was stablished to distinguish between standard and wrong demonstrative sports movements. The proposed method was proved effective and superior through experiments. The research results provide a good reference for the application of 3D image analysis in the recognition of other body behaviors and movements.

1. INTRODUCTION

With its positive effects on health, sports provide an important way to improve the physical quality of the entire population. However, physical education (PE) often faces such problems as nonstandard technical movements or wrong demonstrative movements, which directly affect the teaching quality, and constrains the physical literacy of students. In severe cases, these movements could damage physical functions [1-4]. Standard sports demonstrative movements need to meet the quality standards in terms of direction, sequence, rhythm, trajectory, space, and time [5-9]. To improve PE teaching and practice, it is important to analyze the technical features of each link in the specific sports movement.

High-speed photography can capture instantaneous movements that cannot be recognized with naked eyes. Therefore, this technology has been widely used to judge the sprint movements in track and field competitions, and assess the quality of artistic gymnastics. Recently, a growing attention has been paid to the three-dimensional (3D) image analysis of human movements [10-12]. Piergiovanni and Ryoo [13] sorted out the precautions for shooting the sports process with video recorder, and suggested taking the following measures before shooting: building a high-precision calibration frame containing 4-8 points with known coordinates, setting up relevant equations by the least squares (LS) method, and solving the 3D coordinates of each calibration point. Ng et al. [14] introduced the infrared camera system platforms that support kinematics analysis, including Vicon Motus, Motion Analysis, and Qualisys. Moreno et al. [15] applied 3D force measurement system to measure the dynamic parameters of sports process; despite its simplicity and high precision, the system has difficulty in acquiring detailed dynamic parameters of human joints and muscle strength.

With the development of sports, it is increasingly difficult to complete some sports movements. Meanwhile, the multi-camera synchronous measurement becomes more and more popular, owing to the progress in sports biomechanics [16-18]. Shamsipour et al. [19] connected multiple sync signal generators to measure the movements in track and field, gymnastics, and ball games, and successfully optimized the amount and accuracy of parameter calculation. Asteriadis and Daras [20] combined force plate, infrared high-speed camera system, and surface electromyography into a sports movement test system, and relied on the system to extract and analyze the biomechanical features of the lower limbs of table tennis players in forehand stroke. Wang et al. [21] constructed a biomechanical test system for entry movements of swimmers; the system, consisting of a kinematics test module, a dynamic test module, and a synchronization module, can accurately evaluate the smooth transition from underwater sliding phase to formal swimming phase.

Many scholars have studied sports kinematics extensively, with the aid of image analysis [22, 23]. Through 3D video analysis, Rezazadegan et al. [24] collected the kinematic images and data of the basic movements of women's floor exercise, and quantified the dynamics and kinematics features of athletes in landing, using a simulation model of the actual movement process. Yadav and Sethi [25] explored the movement features, laws, and characteristics of the triple jumps of single figure skaters, and discussed the correlation of movement standardization with the coordination between upper and lower limbs, movement time, center of gravity, as well as the angles and speeds of upper and lower limbs.

Taking the human body as a 3D skeleton of joints connected by rigid bones, every sports movement can be regarded as a skeletal system movement, which is highly robust to changes in illumination, scale, and perspective. This paper aims to make a scientific analysis of the technical features of demonstrative sports movements, and realize accurate evaluation of movement standardization. With the help of 3D image analysis, this paper draws on the current results on skeletal movement recognition, and proposes a method to recognize the standard and wrong demonstrative sports movements, based on 3D convolutional neural network (CNN) and graph theory.

The remainder of this paper is organized as follows: Section 2 presents a 3D posture perception strategy for demonstrative sports movements based on video sequence, and explains the restoration of demonstrator's 3D posture from the video sequence of sports movements; Section 3 details the framework of the proposed recognition system for standard and wrong demonstrative sports movements; Section 4 establishes a 3D CNN that differentiates between standard and wrong demonstrative sports movements; Section 5 verifies the proposed method through experiments; Section 6 puts forward the conclusions.

2. POSTURE PERCEPTION

This paper designs a 3D posture perception strategy that outputs a 3D posture sequence from a video sequence containing demonstrative sports movements. The proposed strategy mainly consists of OpenPose-based two-dimensional (2D) posture perception, 3D posture perception of demonstrative sports movements based on coding network and three-layer decoding neural network, and demonstrator posture restoration based on the correlation and continuity between video frames.

After being compared with DeepCut and Mask R-CNN, the open-source bottom-up OpenPose was chosen as the 2D pose extractor. The input of the extractor is made up of a color image and a central response graph. The loss function can be expressed as:

$$Loss_X^t = \sum_{m=1}^{M}\sum_{x\in X}\left\|H_k^t(x) - H_k^*(x)\right\|_2^2 \tag{1}$$

where, X is the response graph space of key points of human skeleton; M is the number of preset key points of human skeleton; t is the serial number of time periods. Figure 1 presents a demonstrative movement of gymnastics and the key points of the demonstrator's skeleton.

(a) (b)

Figure 1. The demonstrative movement of gymnastics (a) and the key points of the demonstrator's skeleton (b)

OpenPose introduces some affinity fields to estimate human postures. The body feature map package is composed of two networks: the affinity fields F of the body, and the response graphs R of all key points. Each affinity field F can be calculated by:

$$F = \begin{cases} \dfrac{a_{k_{2,i}} - a_{k_{1,i}}}{\left\|a_{k_{2,i}} - a_{k_{1,i}}\right\|_2}, & Yes \\ 0, & Otherwise \end{cases} \tag{2}$$

where, a is the coordinates of a key points of human skeleton; $k_{1,i}$ and $k_{2,i}$ are the key points at the two ends of part b of demonstrator i, respectively. Whether pixel p of a video frame falls on part l of demonstrator i can be judged by:

$$\begin{aligned} &0 \le \frac{a_{k_{2,i}} - a_{k_{1,i}}}{\left\|a_{k_{2,i}} - a_{k_{1,i}}\right\|_2} \cdot \left(p - a_{k_{1,i}}\right) \le L, \\ &\left| \frac{a_{k_{2,i}} - a_{k_{1,i}}}{\left\|a_{k_{2,i}} - a_{k_{1,i}}\right\|_2} \perp \cdot \left(p - a_{k_{1,i}}\right) \right| \le \eta \end{aligned} \tag{3}$$

where, L and η are the length and width threshold of a part of the demonstrator, respectively. The loss function based on some affinity fields can be calculated by:

$$Loss_A^t = \sum_{b=1}^{B}\sum_{x\in X}\left\|A_t^b(x) - F(x)\right\|_2^2 \tag{4}$$

It is a non-deterministic polynomial-time (NP) hard problem to solve the edges between all key points of demonstrator skeleton according to the obtained affinity fields A and the response graphs S of all key points. Therefore, two slack conditions were added: setting up the demonstrator skeleton with the spanning tree of the smallest edge, and independent calculation of the matching between adjacent edges.

From the obtained 2D postures and demonstrator height, a 3D CNN can be constructed to solve the parameters of skinned multi-person linear (SMPL) model $SMPL(\hat{s},\hat{e},\psi): R^{|\hat{s}|\times|\hat{e}|} \rightarrow R^{3N}$, where $\psi=[Q,W,P,O,K]$ is the set of parameters to be obtained by solving the model. Specifically, Q and W are the average model and mixed weight of the demonstrator body, respectively; P and O are the orthogonal basis of the principal shape components and the regression matrix of key points, respectively; $K=[K_1, K_2, \ldots, K_{9M}] \in R^{3N\times 9M}$ is the mixed deformation vector of sports postures. Let ρ be the global rotation coefficient, ε be the translation vector, and υ be the scale factor. Then, it is possible to derive the parameters $\Gamma=[\hat{s},\hat{e},\rho,\varepsilon,\upsilon]$ to be solved for restoring the 3D postures of the demonstrator. The global loss function can be calculated by:

$$\begin{aligned} Loss = {} & \omega_{rep} Loss_{rep} + 1_{3d}\,\omega_{3d} Loss_{3d} \\ & + \omega_{adv} Loss_{adv} + 1_h\,\omega_h Loss_h \end{aligned} \tag{5}$$

where, ω_{rep}, ω_{3d}, ω_{adv}, and ω_h are the weight coefficients of the loss function; $Loss_{3d}$, and $Loss_h$ are binary functions indicating the loss; $Loss_{rep}$ is the deviation of the coordinates a' of key points in 3D demonstrator skeleton projected onto the 2D

plane from the actual coordinates a of the key points in 2D demonstrator skeleton:

$$Loss_{rep} = \sum_i \left\| h_i \left(a_i - \hat{a}_i \right) \right\|_1 \tag{6}$$

where, h_i is a binary function reflecting whether key point i is visible. The projection of coordinates a' can be detailed by:

$$a' = \upsilon \Phi \left(\rho Y \left(\hat{s}, \hat{e} \right) \right) + \varepsilon \tag{7}$$

where, Φ is the orthogonal projection; $Y(\hat{s},\hat{e})$ is the set of key points in 3D demonstrator skeleton obtained through SMPL model. The actual set of key points $(\hat{s},\hat{e})$ in 3D demonstrator skeleton can be computed by Mosh method.

Under projection constraints, the projected coordinates of key points in 3D demonstrator skeleton are basically consistent with those of the key points in 2D demonstrator skeleton. But the non-uniqueness of solutions brings the risk that the generated 3D demonstrator body is not the actual body. To prevent this risk, 3D constraints can be introduced:

$$Loss_{3d} = Loss_{3d-j} + 1_s Loss_{3d-s} \tag{8}$$

$$Loss_{3d-j} = \sum_i \left\| Y_i - \hat{Y}_i \right\|_2^2 \tag{9}$$

$$L_{3d-s} = \sum_i \left\| \left[\hat{s}_i, \hat{s}_i \right] - \left[\hat{s}_i^*, \hat{s}_i^* \right] \right\|_2^2 \tag{10}$$

where, 1_s is a binary function reflecting the loss. If SMPL parameters do not exist, 1_s=0; otherwise, 1_s=1.

Without any constraint, the generated 3D model may have abnormalities, such as crossed limbs or deformed limbs. To enable the parameters Γ of the 3D demonstrator skeleton generated by generative adversarial network CN to approximate the actual distribution, the data-driven rules were configured as discriminant networks DNs, whose loss function can be calculated by:

$$Loss_{adv} = Loss_{CN} + Loss_{DN} \tag{11}$$

$$Loss_{CN} = \sum_i \left(DN_i \left(CN \left(P \right) \right) - 1 \right)^2 \tag{12}$$

$$Loss_{DN} = \left(DN_i \left(\Gamma^* \right) - 1 \right)^2 + DN_i \left(CN \left(P \right) \right)^2 \tag{13}$$

where, i is the serial number of a DN; P is the input video images; Γ^* is the actual parameters of the demonstrator; $Loss_{CN}$ is the loss function of CN aiming to control the output of parameters Γ in DN close to 1; $Loss_{DN}$ is the loss function of DN aiming to enhance the ability to differentiate between actual and generated parameters of the demonstrator.

Considering the accuracy loss in the 3D body restoration from 2D images, a height loss function can be set as:

$$Loss_{HEI} = \left\| g_{HEI} \left(\Gamma \right) - HEI^* \right\|_2^2 \tag{14}$$

where, g_{HEI} is the arithmetic function used to calculate the body height with the parameters of 3D demonstrator skeleton; HEI^* is the actual parameters of the demonstrator skeleton.

To generate a jitter-free and smooth 3D posture sequence of the demonstrator, this paper constructs a posture restoration model based on the 3D postures in the stream of video frames. The model is mainly made up of a known OpenPose network, a coding network, a decoding network (i.e. the iterative regression network in the CN), in addition to input and output images. The posture restoration process of the model is explained in Figure 2. Firstly, the key points in 2D skeleton were extracted by OpenPose from each video frame. The extraction error can be expressed as:

$$Loss_{2d} = \sum_j \sum_k b_{j,k} \left\| \hat{z}_{j,k} - \Phi \left[f_k \left(g_d \left(\hat{v}_j \right) \right) \right] \right\|_1 \tag{15}$$

where, j is the serial number of each frame; k is the serial number of each key point; g_d is the operation function of the decoding network for 2018-dimensional feature space; f is the prediction function of the key points in demonstrator skeleton.

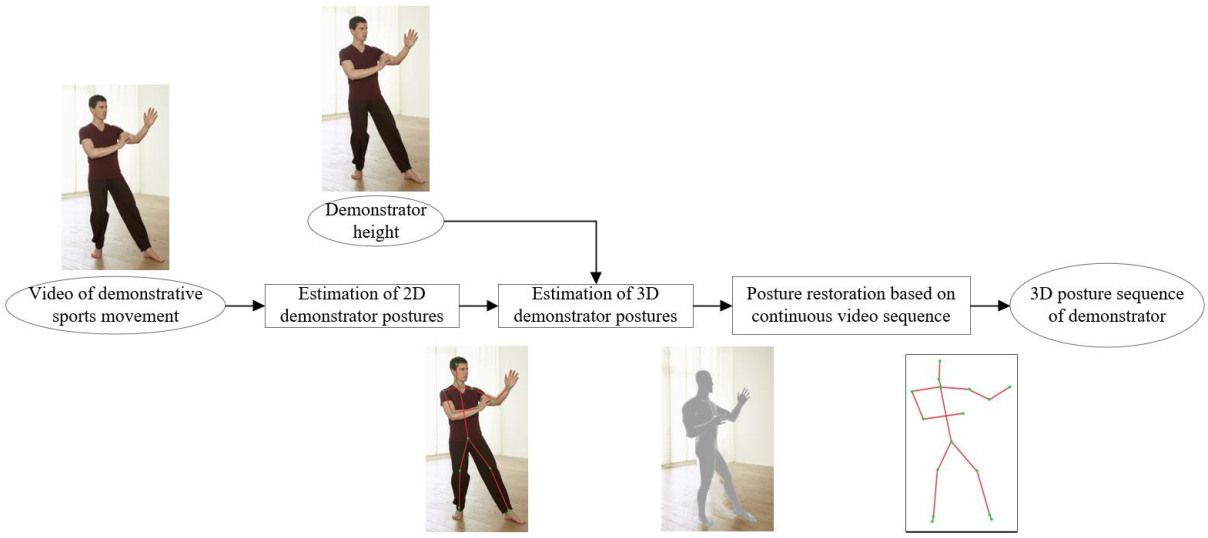


Figure 2. The posture restoration process

The above 2D constraint further optimizes the feature space, producing a more accurate 3D model of demonstrator body. Then, an initial error constraint was introduced to preserve as much of the information that favors the restoration of the initial state as possible:

$$Loss_{3d-INI} = \sum_j \omega_j \left(\left\| \hat{\alpha} - \left[g_d\left(\hat{v}_j\right) \right]_\alpha \right\|_2^2 \right) \tag{16}$$

where, α is the transformed rotation matrix, i.e. the postures of the initially imported 3D demonstrator skeleton; ω_j is the weight coefficient used to preserve the favorable information:

$$\omega_t = \exp\left(-\sum_k b_{j,k} \left\| \hat{z}_{j,k} - \Phi f_k\left(g_d\left(\hat{v}_t\right)\right) \right\|_2^2 \right) \tag{17}$$

The effectiveness of the eigenvector can be measured by the extraction error of 2D key points. The greater the error, the smaller the weight, and the larger the deviation of the decoded feature space from the original state. The measurement can be constrained by the error between the key points in the 3D skeletons of adjacent frames:

$$Loss_{SMO} = \sum_j \sum_k \left\| f_k\left(f_d\left(\hat{v}_j\right)\right) - f_k\left(g_d\left(\hat{v}_{j+1}\right)\right) \right\|_2^2 \tag{18}$$

To sum up, the global constraint function can be defined as:

$$\begin{aligned} Loss_{\text{total}} &= \omega_{2d} Loss_{2d} + \omega_{3d-INI} Loss_{3d-INI} \\ &+ \omega_{SMO} Loss_{SMO} + 1_{HEI}\, w_{HEI}\, loss_{HEI} \end{aligned} \tag{19}$$

3. FRAMEWORK OF RECOGNITION SYSTEM

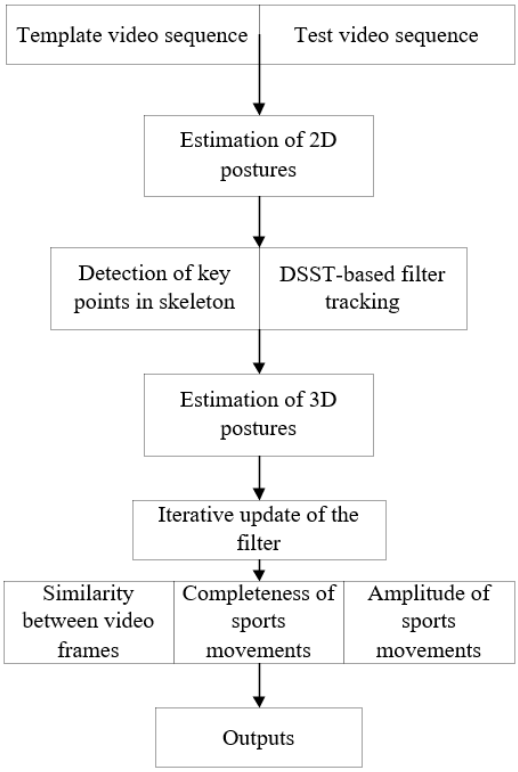


Figure 3. The framework of recognition system
Note: DSST is short for discriminative scale space tracker.

Figure 3 illustrates the framework of the proposed recognition system for standard and wrong demonstrative sports movements. The software program of the system mainly involves frontend, backend, and algorithm modules. The designed system can be implemented in the following steps:

Step 1. Estimate the postures of the demonstrator. Read the template and test video sequences by OpenCV. Import all files as the inputs of the 3D CNN. Calculate the coordinates of skeleton points in the video frames with OpenPose network.

Step 2. Initialize DSST algorithm based on the positions of the demonstrator in the initial frame. Optimize the square error of the collected training samples by:

$$E = \sum_{k=1}^{j} \left\| T_j \cdot r_k - G_k \right\|^2 = \frac{1}{PQ} \left\| \bar{T}^*_j \cdot r^*_k - G^*_k \right\|^2 \tag{20}$$

where, r_k is a training sample; G_k and T_k are the Gaussian feature map and the needed filter, both of which are of the size $P*Q$; r^*_k, G^*_k, and T^*_k are the parameter values after discrete Fourier transform; $\bar{T}^*_k$ is the complex conjugate form of T^*_k. Formula (20) can be minimized by:

$$T^*_k = \frac{\sum_{k=1}^{j} \bar{G}^*_k r^*_k}{\sum_{k=1}^{j} \bar{r}^*_k r^*_k} \tag{21}$$

Iteratively optimize the numerator and denominator in formula (21) separately. After the training, perform discrete Fourier transform on region Δq of the new video frame. Then, calculate the response score of this region:

$$RS = Fourier^{-1}\left\{ \bar{T}_k \cdot \Delta q \right\} \tag{22}$$

Find the maximum response score to track the position of the demonstrator. During the estimation of demonstrator scale, calculate the loss function that considers position and scale dimensions:

$$Loss = \left\| \sum_{l=1}^{L} T^l \cdot r^l - G \right\|^2 + \zeta \sum_{l=1}^{d} \left\| T^l \right\|^2 \tag{23}$$

where, L is the scale dimension; ζ is the regularization term. Solve the above formula to obtain H in the Fourier space:

$$T^{*l}_k = \frac{\bar{G} r^{*l}_k}{\sum_{t=1}^{d} r^{*t} r^{*\bar{t}} + \zeta} \tag{24}$$

To improve the robustness of the tracking algorithm, denote the iterative numerator and denominator in formula (24) as δ_j and μ_j, respectively. Express the update strategies of δ_j and μ_j as:

$$\delta^l_j = (1-\lambda)\delta^l_{j-1} + \lambda G_k \bar{r}^l_j \tag{25}$$

$$\mu_t = (1-\lambda)\mu_{j-1} + \lambda \sum_{t=1}^{d} r^{*t}_k r^{*\bar{t}}_k \tag{26}$$

Next, calculate response score that considers the scale dimension:

$$RS' = Fourier^{-1}\left\{\frac{\sum_{l=1}^{L}\bar{\delta}^{l}\Delta q^{l}}{\mu+\zeta}\right\} \tag{27}$$

Step 3. To measure the similarity between video frames, design the following angles and extract the angle features of the demonstrator movements: wrist-elbow-shoulder, head-neck-shoulder, neck-shoulder-elbow, shoulder-hip-knee, and hip-knee-ankle.

Step 4. Conduct pairwise comparison between adjacent frames to complete the detection of key frames. Record and save the time and frame number of the current template video frame. Based on the angle features obtained in Step 3, calculate the completeness and amplitude of demonstrative movements, and the similarity between video frames. Finally, output the calculated results.

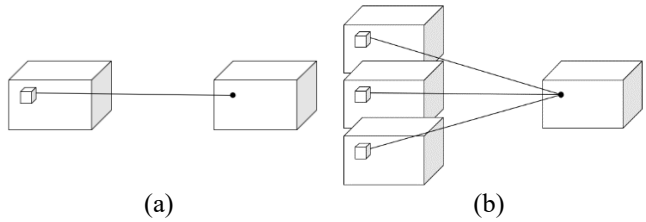
4. CONSTRUCTION OF 3D CNN

This paper processes video frames with 3D CNN, which adds the time dimension to the 2D CNN. In the 3D CNN, the multiple continuous video frames of demonstrative sports movements are superimposed through 3D convolutions. 2D convolution is compared with 3D convolution in Figure 4.

To obtain ideal information about demonstrative sports movements, the feature map of each convolution layer was connected to the inputs of multiple adjacent frames. Let ε_{ij} be the error of feature map j of convolution layer i; R_i, S_i, and T_i be the space-time dimension of the 3D kernel of convolution layer i. Then, the coordinates (a,b,c) of pixel i of feature map j of convolution layer i can be expressed as:

$$v_{ij}^{xyz} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \tag{28}$$

where, tanh is the hyperbolic tangent function; ω_{ijk}^{rst} is the kernel weight matrix of feature map k of convolution layer i-1.



(a) (b)

Figure 4. The comparison between (a) 2D convolution and (b) 3D convolution

The eigenvector extracted by the current 3D CNN often faces the problem of information loss, owing to the neglection of high-level movement information. The solve the problem, this paper proposes an improved 3D deep CNN for the recognition of demonstrative sports movements (Figure 5).

As shown in Figure 5, the first layer of the improved 3D deep CNN adopts five channels: x-direction gradient, x-direction optical flow, y-direction gradient, y-direction optical flow, and grayscale. This layer is followed by five convolution layers, five pooling layers, and two fully-connected layers. The output layer is a softmax classifier.

The standard and wrong demonstrative sports movements include 121 different gymnastics movements. Eight consecutive video frames of the size 256×171 were randomly selected as network inputs. Thus, the size of the input frame is 5×8×256×171 in size. In the first to fifth convolution layers, the number of kernels is 128, 256, 256, 256, and 512 in turn; all the kernels are of the size 3×3×3. In the pooling layers, the pooling kernels are of the size 2×2×2. The output signal of each pooling layer are eight times smaller than the corresponding input signal. The two fully-connected layers output a 2,048-dimensional eigenvector, which is processed by the softmax classifier to obtain the final predicted classes.

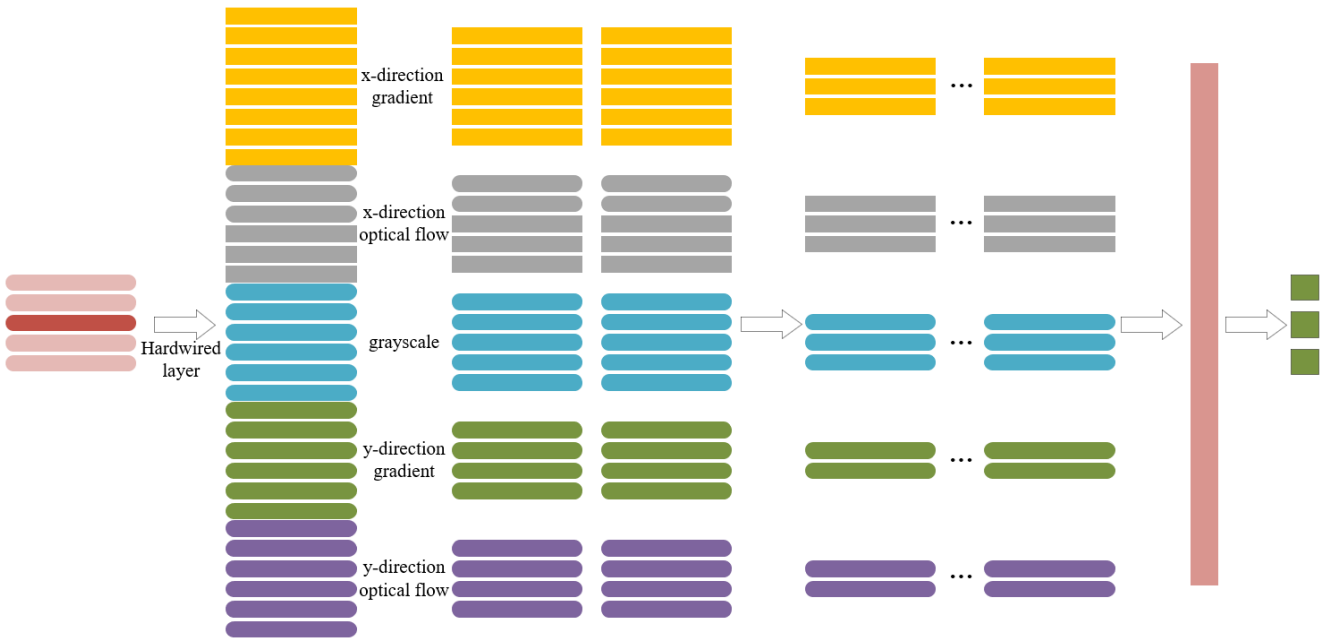


Figure 5. The structure of the 3D deep CNN

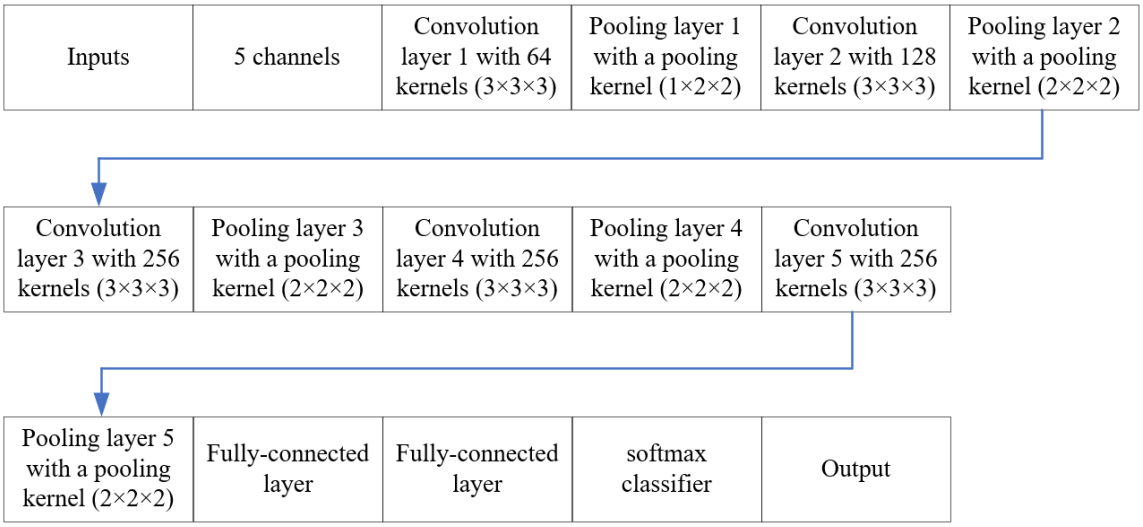


Figure 6. The structure of the improved 3D deep CNN

Figure 6 sums up the structure of the proposed 3D deep CNN.

To improve the recognition accuracy of standard and wrong demonstrative movements in complex sports (e.g. gymnastics), the number of continuous video frames was increased to above 16. Hence, an auxiliary feature regularization module was fused into the 3D deep CNN. Based on the historical images of demonstrative sports movements, the change of a pixel in a period was deduced to represent a movement in the form of image brightness. Let Δd and φ be the decay parameter and update function of the brightness at a pixel in historical images, respectively. Then, the brightness intensity IV of that pixel can be calculated by:

$$IV_{\sigma}(a,b,t)=\begin{cases}\sigma, \text{if}\,\phi(a,b,t)=1\\ \max\left(0, IV_{\sigma}(a,b,t)-\Delta d\right), otherwise\end{cases} \tag{29}$$

where, (a,b) is the position of the pixel; t is the current time; σ is the period of the continuous frames. The commonly used frame difference method can be expressed by:

$$\phi(a,b,t)=\begin{cases}1, if\ \left|IV(a,b,t)-IV\left((a,b,t)\pm e\right)\right|\geq\beta\\ 0, otherwise\end{cases} \tag{30}$$

where, β is a preset difference threshold related to video scenario; $IV(a,b,t)$ is the brightness intensity of pixel (a,b) in frame t; e is the distance between video frames.

The appearance, movement shape, and movement information were preserved by the scale invariant feature transform (SIFT) descriptor, and the historical images on movement edges. In the designed network, the last layer is a hidden layer, which receives auxiliary features and outputs the result via auxiliary output points. Through the regularization of auxiliary rules, the extracted standard and wrong movements can approximate the output of auxiliary features.

5. EXPERIMENTS AND RESULT ANALYSIS

First, a 3D deep CNN was constructed based on the structural information in Figure 6. The recognition effect depends heavily on kernel size. It is important to select a rational kernel size, which balances the computing complexity with information enhancement. For this purpose, contrastive experiments were conducted to compare the recognition accuracies of demonstrative sports movements with kernels of different sizes (Figure 7).

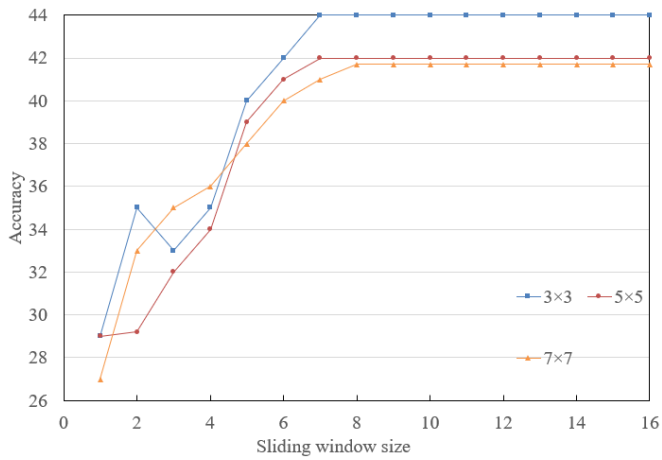


Figure 7. The recognition accuracies with kernels of different sizes

As shown in Figure 7, the standard and wrong demonstrative sports movements were recognized most effectively under the time dimension of three. Therefore, the 3D kernels of the size 3×3×3 were selected for the proposed 3D deep CNN.

To prevent early occurrence of overfitting, the pooling kernel of the first pooling layer was set to 1×2×2, and that of the other pooling layers was set to 2×2×2, based on the outputs of the previous convolution layer. This kernel setting guarantees the pooling effect, without losing any information on complex sports movements.

To verify its performance, the proposed 3D deep CNN was compared with LSTM and 2D CNN. As shown in Figure 8, our network achieved more desirable results than the two contrastive methods.

Next, the proposed network was applied to extract the angle features of demonstrator movements, including wrist-elbow-shoulder, head-neck-shoulder, neck-shoulder-elbow, shoulder-hip-knee, and hip-knee-ankle.

Table 1 compares the key points of demonstrator skeleton in standard movements and test movements, and provides the angles, and angular velocities of the horizontal axes of shoulder and hip, as well as the time consumption of movements in the template and test video sequences.

As shown in Table 1, the template and test video sequences differed slightly in time consumption, and the angles of the horizontal axes of shoulder and hip. However, the angular velocities of shoulder and hip in test video sequence were much slower than those in template video sequence. This means the demonstrative sports movements in test video sequence are slow, and substandard. Although the movements have no error, the angles of the joints should be increased reasonably.

To verify its recognition effect on standard and wrong demonstrative sports movements, the proposed 3D deep CNN was compared with BPNN, LSTM, fuzzy BPNN, GA-PSO optimized BPNN, 2D CNN, and visual adaptive LSTM. As shown in Table 2, our model achieved better recognition results than all the other methods in the recognition of standard and wrong demonstrative sports movements. The recognition accuracies on standard and wrong movements of our method were 1.3% and 1.5% higher than those of visual adaptive LSTM, the best performing method among the contrastive methods. The results of our method were more universal and clearer than those of any other neural network, which can be applied in actual scenarios effectively and conveniently.

Furthermore, our method was compared with fuzzy BPNN, 2D CNN, and visual adaptive LSTM under upper and lower limb movements of 5 demonstrators. The recognition effect of each method was measured by accuracy, recall, and F-score. As shown in Table 3, our method realized balanced recognition effects on upper and lower limb movements, as measured by the three metrics, and controlled the difference between the recognition effects on the two kinds of movements to a small level. The results further confirm the superiority of our method.

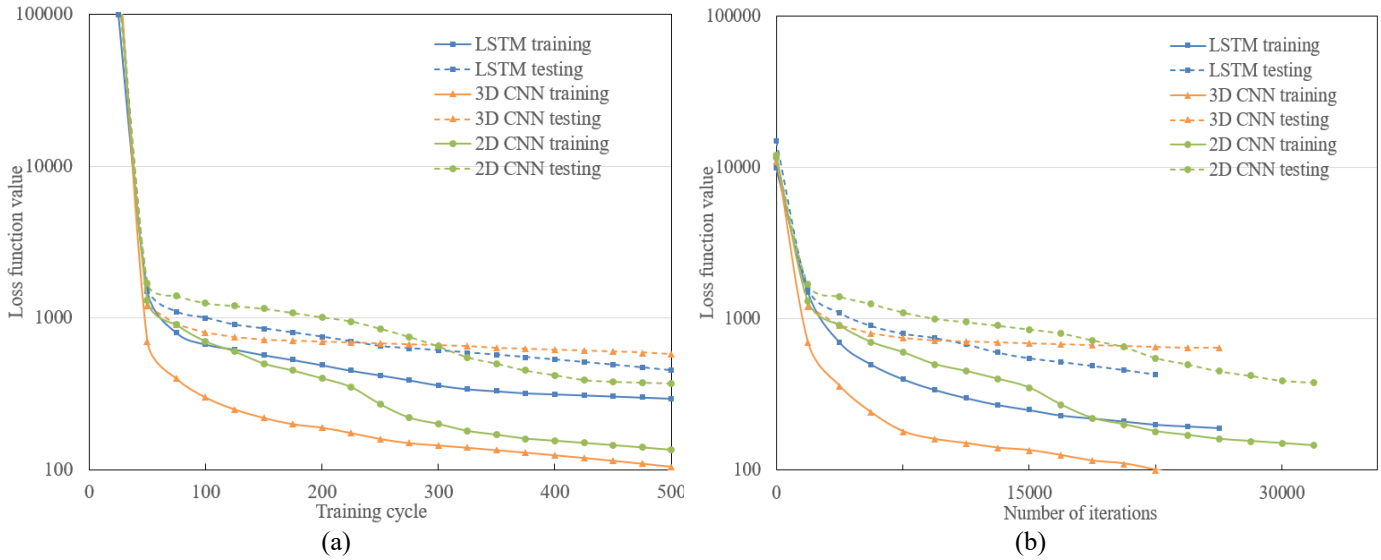


Figure 8. The comparison of training performance (a) and convergence (b)
Note: LSTM is short for long short-term memory.

Table 1. The comparison of key points between standard and test movements

Type	Angle of shoulder	Angle of hip	Angular velocity of shoulder	Angular velocity of hip	Time consumption
Template	263.363	121.609	321.650	242.124	0.842
Test	246.723	119.346	241.199	119.887	0.897

Table 2. The comparison of recognition accuracies on standard and wrong movements

Method	Recognition accuracy	
	Standard movements	Wrong movements
BPNN	78.7	78.9
LSTM	82.2	82.6
Fuzzy BPNN	83.8	88.6
GA-PSO optimized BPNN	87.9	88.2
2D CNN	88.3	89.2
Visual adaptive LSTM	90.1	90.9
3D deep CNN	91.2	92.3

Note: BPNN is short for backpropagation neural network; GA-PSO is short for genetic algorithm and particle swarm optimization.

Table 3. The comparison of recognition performance on upper and lower limb movements

		Fuzzy BPNN		2D CNN	
		Upper limb movements	Lower limb movements	Upper limb movements	Lower limb movements
Demonstrator 1	**Precision**	88.91	88.00	85.12	86.09
	Recall	87.21	87.65	89.67	86.53
	F-score	87.89	86.47	87.49	86.92
Demonstrator 2	**Precision**	84.00	85.98	87.48	87.29
	Recall	85.63	86.74	87.89	88.72
	F-score	87.68	88.43	87.59	88.52
Demonstrator 3	**Precision**	86.25	87.68	86.44	85.94
	Recall	89.86	88.59	86.92	88.22
	F-score	87.21	85.67	86.11	79.57
Demonstrator 4	**Precision**	80.91	84.47	83.02	84.70
	Recall	82.89	82.87	84.80	85.24
	F-score	83.99	82.86	86.92	87.11
Demonstrator 5	**Precision**	88.02	85.15	85.26	86.17
	Recall	87.31	87.57	88.39	90.11
	F-score	82.34	83.91	84.65	86.94
		Visual adaptive LSTM		Our model	
		Upper limb movements	Lower limb movements	Upper limb movements	Lower limb movements
Demonstrator 1	**Precision**	87.35	88.92	90.75	91.68
	Recall	87.77	88.57	89.79	91.93
	F-score	88.32	89.92	91.96	93.24
Demonstrator 2	**Precision**	88.35	89.10	93.12	90.06
	Recall	85.08	89.64	91.64	92.43
	F-score	89.19	88.17	90.48	92.05
Demonstrator 3	**Precision**	87.24	87.14	89.87	88.09
	Recall	91.45	89.43	92.98	92.16
	F-score	82.37	85.79	88.21	88.89
Demonstrator 4	**Precision**	85.01	86.96	89.07	89.56
	Recall	85.96	86.91	87.71	89.80
	F-score	88.77	88.98	90.48	91.72
Demonstrator 5	**Precision**	87.76	87.26	89.71	90.35
	Recall	89.67	89.87	90.75	91.73
	F-score	89.29	89.93	91.27	92.59

6. CONCLUSIONS

Drawing on 3D image analysis, this paper proposes a method to recognize the standard and wrong demonstrative sports movements, based on a self-designed 3D convolutional neural network (CNN) and graph theory. Firstly, a 3D posture perception strategy was constructed for demonstrative sports movements based on video sequence. Next, the authors introduced the framework of the recognition system for standard and wrong demonstrative sports movements. After that, a 3D CNN was stablished to distinguish between standard and wrong demonstrative sports movements. Experimental results demonstrate that our method outshined the other methods in training cycle and convergence time, output more universal and clearer recognition results than the other methods on the standard and wrong movements, and achieved balanced results on accuracy, recall, and F-score. To sum up, this paper develops an advantageous model for the recognition of standard and wrong demonstrative sports movements.

ACKNOWLEDGEMENTS

This paper was supported by the Science Research Plan Project of Shaanxi Province Education Department (Grant No.: 18JK0025) and the Key Subsidizing Item of Scientific Research of Baoji University of Arts and Science (Grant No.: ZK2017005).

REFERENCES

[1] Ijjina, E.P. (2020). Action recognition in sports videos using stacked auto encoder and HOG3D features. Proceedings of the Third International Conference on Computational Intelligence and Informatics, Singapore, pp. 849-856. https://doi.org/10.1007/978-981-15-1480-7_79

[2] Ichige, R., Aoki, Y. (2020). Action recognition in sports video considering location information. International Workshop on Frontiers of Computer Vision, Singapore, pp. 150-164. https://doi.org/10.1007/978-981-15-4818-5_12

[3] Kong, L., Huang, D., Qin, J., Wang, Y. (2019). A joint framework for athlete tracking and action recognition in sports videos. IEEE Transactions on Circuits and Systems for Video Technology, 30(2): 532-548. https://doi.org/10.1109/TCSVT.2019.2893318

[4] Zhou, E., Zhang, H. (2020). Human action recognition towards massive-scale sport sceneries based on deep multi-model feature fusion. Signal Processing: Image Communication, 84: 115802. https://doi.org/10.1016/j.image.2020.115802

[5] Tejero-de-Pablos, A., Nakashima, Y., Sato, T., Yokoya, N., Linna, M., Rahtu, E. (2018). Summarization of user-generated sports video by using deep action recognition features. IEEE Transactions on Multimedia, 20(8): 2000-2011. https://doi.org/10.1109/TMM.2018.2794265

[6] Martin, P. E., Benois-Pineau, J., Péteri, R., & Morlier, J. (2018). Sport action recognition with Siamese spatio-temporal CNNs: Application to table tennis. 2018 International Conference on Content-Based Multimedia Indexing (CBMI), La Rochelle, pp. 1-6. https://doi.org/10.1109/CBMI.2018.8516488

[7] Tejero-de-Pablos, A., Nakashima, Y., Sato, T., Yokoya, N. (2016). Human action recognition-based video summarization for RGB-D personal sports video. 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, pp. 1-6. https://doi.org/10.1109/ICME.2016.7552938

[8] Liu, G., Zhang, D., Li, H. (2014). Research on action recognition of player in broadcast sports video. International Journal of Multimedia and Ubiquitous Engineering, 9(10): 297-306.

[9] Mishra, O., Kavimandan, P.S., Tripathi, M.M., Kapoor, R., Yadav, K. (2020). Human action recognition using a new hybrid descriptor. Advances in VLSI, Communication, and Signal Processing, Singapore, pp. 527-536. https://doi.org/10.1007/978-981-15-6840-4_43

[10] Hussein, N., Gavves, E., Smeulders, A.W. (2019). Timeception for complex action recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 254-263. https://doi.org/10.1109/CVPR.2019.00034

[11] Sanou, I., Conte, D., Cardot, H. (2019). An extensible deep architecture for action recognition problem. In Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019), pp. 191-199. https://doi.org/10.5220/0007253301910199

[12] Papadopoulos, K., Demisse, G., Ghorbel, E., Antunes, M., Aouada, D., Ottersten, B. (2019). Localized trajectories for 2d and 3d action recognition. Sensors, 19(16): 3503. https://doi.org/10.3390/s19163503

[13] Piergiovanni, A.J., Ryoo, M.S. (2019). Representation flow for action recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 9937-9945. https://doi.org/10.1109/CVPR.2019.01018

[14] Ng, J.Y.H., Choi, J., Neumann, J., Davis, L.S. (2018). Actionflownet: Learning motion representation for action recognition. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, pp. 1616-1624. https://doi.org/10.1109/WACV.2018.00179

[15] Moreno, W., Garzón, G., Martínez, F. (2018). Frame-level covariance descriptor for action recognition. In Colombian Conference on Computing, Cartagena, Colombia, pp. 276-290. https://doi.org/10.1007/978-3-319-98998-3_22

[16] Papadopoulos, G.T., Daras, P. (2016). Human action recognition using 3d reconstruction data. IEEE Transactions on Circuits and Systems for Video Technology, 28(8): 1807-1823. https://doi.org/10.1109/TCSVT.2016.2643161

[17] Michalczuk, A., Wereszczyński, K., Segen, J., Josiński, H., Wojciechowski, K., Bąk, A., Wojciechowski, S., Drabik, A., Kulbacki, M. (2017). Manifold methods for action recognition. In Asian Conference on Intelligent Information and Database Systems, Kanazawa, Japan, pp. 613-622. Springer, Cham. https://doi.org/10.1007/978-3-319-54430-4_59

[18] Fernando, B., Gavves, E., Oramas, J., Ghodrati, A., Tuytelaars, T. (2016). Rank pooling for action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4): 773-787. https://doi.org/10.1109/TPAMI.2016.2558148

[19] Shamsipour, G., Shanbehzadeh, J., Sarrafzadeh, H. (2017). Human action recognition by conceptual features. Proceedings of the International Multi Conference of Engineers and Computer Scientists 2017 Vol I, IMECS 2017, Hong Kong.

[20] Asteriadis, S., Daras, P. (2017). Landmark-based multimodal human action recognition. Multimedia Tools and Applications, 76(3): 4505-4521. https://doi.org/10.1007/s11042-016-3945-6

[21] Wang, L., Li, R., Fang, Y. (2017). Power difference template for action recognition. Machine Vision and Applications, 28(5-6): 463-473. https://doi.org/10.1007/s00138-017-0848-0

[22] Singh, S., Arora, C., Jawahar, C.V. (2017). Trajectory aligned features for first person action recognition. Pattern Recognition, 62: 45-55. https://doi.org/10.1016/j.patcog.2016.07.031

[23] Anwer, R.M., Khan, F.S., van de Weijer, J., Laaksonen, J. (2017). Top-down deep appearance attention for action recognition. Scandinavian Conference on Image Analysis, Tromsø, Norway, pp. 297-309. https://doi.org/10.1007/978-3-319-59126-1_25

[24] Rezazadegan, F., Shirazi, S., Upcrofit, B., Milford, M. (2017). Action recognition: from static datasets to moving robots. 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, pp. 3185-3191. https://doi.org/10.1109/ICRA.2017.7989361

[25] Yadav, G.K., Sethi, A. (2017). Action recognition using spatio-temporal differential motion. 2017 IEEE International Conference on Image Processing (ICIP), Beijing, pp. 3415-3419. https://doi.org/10.1109/ICIP.2017.8296916