Hindawi
Scientific Programming
Volume 2022, Article ID 7349548, 11 pages
https://doi.org/10.1155/2022/7349548

*Research Article*

# Motion Action Analysis at Basketball Sports Scene Based on Image Processing

**Jun Liu**

*School of Physical Education, University of South China, Hengyang 421001, China*

Correspondence should be addressed to Jun Liu; 2001001439@usc.edu.cn

Received 9 December 2021; Revised 17 January 2022; Accepted 11 February 2022; Published 7 March 2022

Academic Editor: Le Sun

To solve the problems of low accuracy and high time cost in manual recording and statistics of basketball data, an automatic analysis method of motion action under the basketball sports scene based on the spatial temporal graph convolutional neural network is proposed. By using the graph structure in the data structure to model the joints and limbs of the human body, and using the spatial temporal graph structure to model the posture action, the extraction and estimation of human body posture in basketball sports scenes are realized. Then, training combined with transfer learning, the recognition of motion fuzzy posture is realized through the classification and application of a label subset. Finally, using the self-made OpenCV to collect and calibrate NBA basketball videos, the effectiveness of the proposed method is verified by analyzing the motion action. The results show that the proposed method based on the spatial temporal graph convolutional neural network can recognize all kinds of movements in different basketball scenes. The average recognition accuracy is more than 75%. It can be seen that the method has certain practical application value. Compared with the common motion analysis method feature descriptors, the motion action analysis method based on the spatial temporal graph convolution neural network has higher identification accuracy and can be used for motion action analysis in the actual basketball sports scenes.

## 1. Related Work

Basketball is one of the most popular sports competitions. The analysis of motion action in the basketball sports scene is helpful to improve basketball players' skills. In addition, it can make basketball coaches and athletes quickly master their own sports characteristics. At present, the analysis mainly relies on manual, and the posture estimation is processed by manual marking basketball video. This method usually has problems of low efficiency, low accuracy, high cost, and so on. To solve the above problems, Yu et al. proposed to use MeanShift to process and track the features of motion videos. The tracking and recognition accuracy of this method is 96.04% and 97.10%, respectively, which has ideal effects [1]. Liu et al. proposed an improved ghosting suppression and adaptive visual background extraction algorithm to effectively remove the ghosting problem in motion videos [2]. Li et al. detected and tracked moving targets by combining FPGA and image processing, which realize the functions such as image acquisition, image gray scale, image filtering, and interframe difference [3]. Bin et al. recognized students' standing behavior in a class based on the region of interest (ROI) and face tracking [4]. Huang detected 3d image targets and introduced a deep learning algorithm, thus greatly improving the accuracy of detection [5]. In addition to the above studies, Sun and Manikandaprabu et al. also proposed target detection and tracking methods [6–12]. The above research provides a lot of useful methods for the tracking of motion targets. Therefore, this paper combines basketball movement to detect and recognize basketball motions so as to provide a new method for the processing of sports video images.

The motion action analysis at the basketball sports scene has made great progress, but its overall performance still needs to be improved. On the one hand, the prediction effect of human posture joint points in the basketball sports scene is not satisfactory. On the other hand, the boundary of estimating motion in basketball is relatively fuzzy, which increases the difficulty of research [13–15]. Therefore, in order to solve the above problems, on the basis of the existing research, utilizing the powerful learning ability of the spatial temporal graph convolutional network (ST-GCN), this study proposes a method of analysis of motion action in basketball sports scene based on image processing and spatial temporal convolutional neural network. What is more, by using the graph structure in the data structure to model the human joint points and limbs, and using the spatial temporal graph structure to model the posture action, the human posture in the basketball sports scene is extracted and estimated. Then, by dividing and applying the label subset, and combining it with migration learning training, the recognition of motion fuzzy posture is realized.

## 2. Introduction of Spatial Temporal Graph Convolutional Neural Networks

The spatial temporal graph convolution neural network redefines convolution according to the graph structure, and it enables the graph structure to perform convolution operations. In 2D image convolution, the feature maps of the whole process are two-dimensional pixels. The convolution step is set as 1, and 0 is added at the appropriate position of boundary to obtain the output feature graph with the same size as the input feature graph. For the input *f* in *c* channels, convolution kernel with size $a * b$ is adopted for convolution; then, the output feature map of position $(x, y)$ is as follows [16]:

$$\begin{aligned} g(x,y) &= f(x,y) * W(x,y) \\ &= \sum_{s=-a}^{a}\sum_{t=-b}^{b} f(s,t)W(s-x,t-y). \end{aligned} \tag{1}$$

In the convolutional neural network, the convolution of the convolution kernel $W$ is the weighted overlay of the corresponding position of the image and the convolution kernel, so the above equation can be rewritten as

$$g(x,y) = \sum_{h=-a}^{a}\sum_{w=-b}^{b} f(p(x,y,h,w)) \cdot W(h,w), \tag{2}$$

where $p$ represents the sampling function, which is responsible for extracting the field of $(x, y)$ and $(x, y)$ itself, which can be expressed as:

$$p(x,y,h,w) = (x,y) + p'(p,w). \tag{3}$$

Here, $W$ is the matrix of $c$ channels, the weighted result obtained from the input sampling inner product of $c$ channels, which represents the weight function.

Formula (2) is extended and graph convolution is defined as follows:

(1) Feature mapping of all nodes (including c-dimensional feature vector) is

$$f_{\text{in}}^{t}: V^{t} \longrightarrow R^{c}. \tag{4}$$

(2) In the image field, the sampling function $p(h,w)$ extracts the points around the center of gravity. In the image structure, for node $v_{ti}$, the sampling function extracts its adjacent point set $B(v_{ti})\{v_{tj}|d(v_{tj},v_{ti}) \le D\}$, where $d(v_{tj},v_{ti})$ represents the minimum distance between the nodes $v_{tj}$ and $v_{ti}$. Therefore, the sampling function $p$: $B(v_{ti}) \longrightarrow V$ can be expressed as

$$p(v_{ti},v_{tj}) = v_{tj}. \tag{5}$$

Considering that the connection between human body joints is sparse, this study takes the joint whose adjacent distance is 1, so $D = 1$ is set.

(3) Two-dimensional image pixels are arranged in squares. Any location field is arranged from top to bottom and from left to right. However, for a general graph structure, adjacent nodes have no fixed order. So instead of labeling and building a weight function node by node, this paper divides the set of the adjacent node $B(v_{ti})$ of the node $v_{ti}$ into a fixed number of $K$ subsets. Meanwhile, it codes the $(c, K)$ dimensional tensor to map adjacent nodes to corresponding label subsets:

$$l_{ti}: B(v_{ti}) \longrightarrow \{0,\ldots,K-1\}. \tag{6}$$

The weight function can be expressed as

$$W(v_{ti},v_{tj}) = W'(l_{ti}(v_{tj})). \tag{7}$$

Using the newly defined sampling function and weight function to rewrite formula (2), then we get

$$f_{\text{out}}(v_{ti}) = \sum_{v_{tj}\in B(v_{ti})} \frac{1}{z_{ti}(v_{tj})} f_{\text{in}}(p(v_{ti},v_{tj})) \cdot W(v_{ti},v_{tj}), \tag{8}$$

where $Z_{ti}(v_{tj})$ is the normalized term, which is equal to the number of subsets. And it is used to measure the influence of different subsets on the output result, which can be calculated by the formula as follows:

$$Z_{ti}(v_{tj}) = \left|\{v_{tk}|l_{ti}(v_{tk}) = l_{ti}(v_{tj})\}\right|. \tag{9}$$

Substituting formulas (5) and (7) into formula (9),we obtain

$$f_{\text{out}}(v_{ti}) = \sum_{v_{tj}\in B(v_{ti})} \frac{1}{z_{ti}(v_{tj})} f_{\text{in}}(v_{tj}) \cdot W'(l_{ti}(v_{tj})). \tag{10}$$

(4) Formula (10) is used to establish the spatial temporal graph convolution model of human posture sequence. First of all, the two adjacent frames with the same node are connected according to the graph structure to form the edge set $E_F$. Then, multiple spatial graphs are connected into the spatial temporal structure, which realizes the spatial temporal graph convolution. Finally, the spatial adjacent point set is extended to adjacent frame nodes as follows [17, 18]:

$$B\left(v_{ti}\right) = \left\{ v_{qj} | d\left(v_{qi}, v_{ti}\right) \le K, |q - t| \le \left| \frac{\Gamma}{2} \right| \right\}. \tag{11}$$

Here, $\Gamma$ is the parameter, representing the time length of the spatial temporal convolution kernel. And it is responsible for setting the distance threshold of adjacent nodes added into the subset to less than $\Gamma/2$ from $v_{ti}$ in the time axle distance.

The spatial temporal convolution sampling function is the same as the convolution sampling function of each frame graph in formula (5). The weight function is for the root node $v_{\text{ti}}$, and the label mapping $l_{\text{ST}}\,(v_{qj})$ of adjacent node set of the spatial temporal graph structure can be expressed as

$$l_{\text{ST}}\left(v_{qi}\right) = l_{ti}\left(v_{tj}\right) + \left(q - t + \left[\frac{\Gamma}{2}\right]\right) \times K. \tag{12}$$

Here, $l_{ti}\,(v_{tj})$ represents the label mapping of the adjacent node set of node $v_{ti}$ in each frame.

## 3. Basketball Motion Analysis Method Based on Spatial Temporal Graph Convolutional Network

*3.1. Overall Process.* According to the characteristics of the above spatial temporal graph convolution network, the specific process of the basketball motion analysis method is designed, as shown in Figure 1. First of all, according to the node sequence formed by each human body joint of input multiple frames, the label subset is divided by the label division strategy. Then, the input tensor is constructed by transforming the spatial temporal graph convolution network into spatial temporal graph convolution. Finally, using the spatial temporal graph convolutional neural network to train and classify output, the analysis of basketball movement is realized. Each key part is explained as follows.

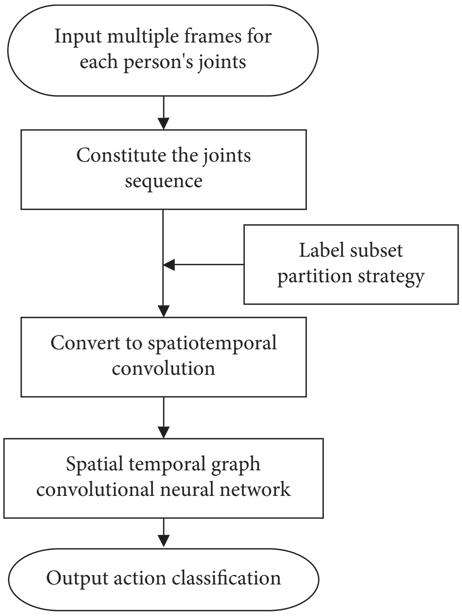

Input multiple frames for each person's joints
Constitute the joints sequence
Label subset partition strategy
Convert to spatiotemporal convolution
Spatial temporal graph convolutional neural network
Output action classification

Figure 1: Basketball movement analysis flow based on spatio-temporal graph convolution network.

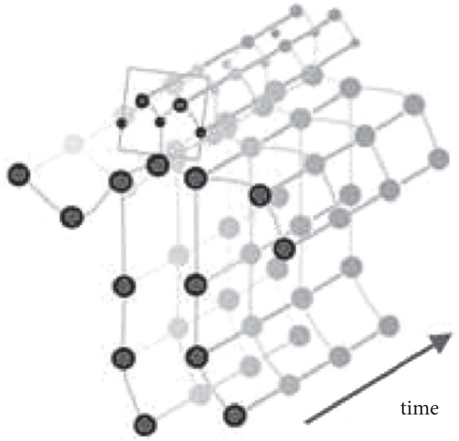


Figure 2: Temporal and spatial diagram of human joint sequence.

*3.2. Construction of the Structural Input of Human Body Joint Sequence Diagram.* According to the multiple joint matching algorithm, the graph structure in the data structure is adopted to model the human body joints and limbs, and the spatial temporal graph structure is adopted to model the posture action, as shown in Figure 2 [19–22].

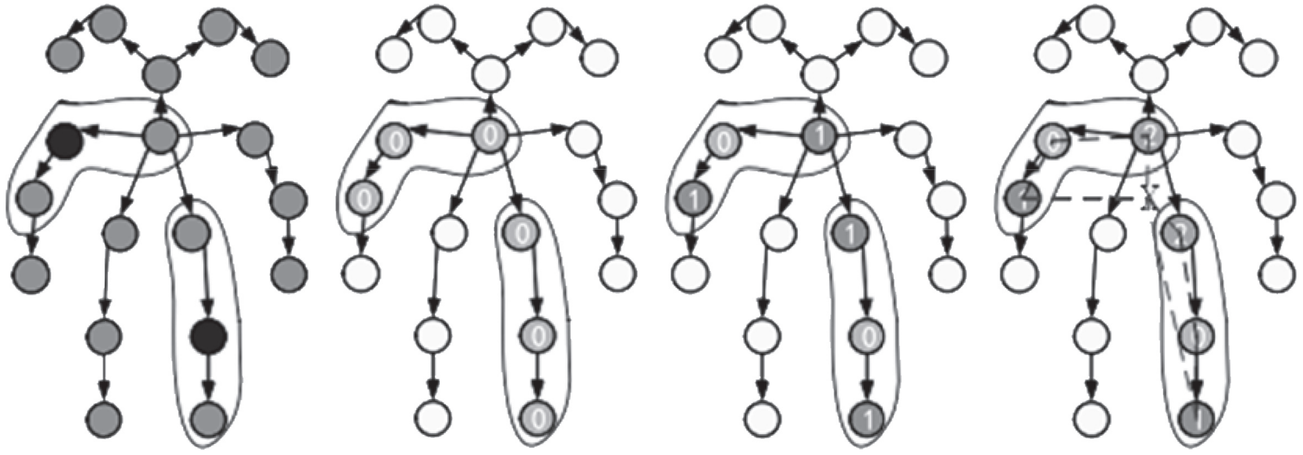
For a $T$ frame, the basketball movement video with $N$ joint posture sequences of each frame can be defined as an undirected temporal and spatial diagram $G = (V, E)$, where $V$ is the input of the convolutional neural network, and it represents the total number of joints in posture sequence [23]. Calculating by formula (13), we obtain the corresponding coordinate confidence of each coordinate point and posture estimation output heat map. In addition, the edge of the spatial temporal graph structure can be decomposed into the edge set of each frame and the edge set between two adjacent frames, which are expressed as formulas (14) and (15), respectively, where $H$ represents the joint of the human limb, and all edges of $E_F$ represent the locus of the joint [24].

$$V = \left\{v_{ti} | t = 1, \ldots, T, i = 1, \ldots, N\right\}. \tag{13}$$

$$E_s = \left\{v_{ti}v_{tj} | t = \tau, (i, j) \in H\right\}, \tag{14}$$

$$E_F = \left\{v_{ti}v_{(t+1)i}\right\}. \tag{15}$$

*3.3. Label Subset Partition Strategy.* The subset of labels in this study is divided by reference to the ST-GCN partition strategy. ST-GCN partition strategy includes unified partition, partition by distance, and partition by spatial structure, as shown in Figure 3. In the figure, Figure 3(b) is a unified partition strategy, which is the most direct and simple partition strategy. By dividing the whole set of adjacent point, the corresponding graph convolution is calculated as the inner product of feature vectors and weight vectors of



(a) (b) (c) (d)

Figure 3: Label subset partition strategy.

each adjacent node to $v_{ij}$. Therefore, it can be seen that the unified partition strategy is to calculate the inner product of all adjacent nodes' average feature vector and weight vector, which is easy to lead to the loss of local features. So, this method is not the best posture sequence classification method [25].

Figure 3(c) is the partition strategy by distance, which is based on the distance $d(\cdot, v_{ti})$ from the root node $v_{ti}$. In this study, $D$ is set to 1, so the root node itself can be regarded as a subset, that is, $D = 0$. The adjacent nodes with distance $D = 1$ can form a subset. Therefore, the partition strategy can include two vectors with different weights to model the local differential characteristics. Label quantity divided by distance is $K = 2$, and the label is

$$l_{ti}(v_{ti}) = d\left(v_{tj}\cdot v_{ti}\right). \quad (16)$$

Figure 3(d) shows the subset partition of adjacent point labels based on the spatial distribution of human body joints, where $X$ is the center of the human body, and the adjacent point labels include three label subsets, such as the root node itself, centrifugal group, and centripetal group. In this paper, the center of gravity of the human body is obtained by averaging the coordinates of all the nodes. According to the spatial distribution, the number of labels is $K = 3$, and the labels are

$$l_{ti}\left(v_{tj}\right) = \begin{cases} 0, & \text{if } r_j = r_i, \\ 1, & \text{if } r_j = r_i, \\ 2, & \text{if } r_j = r_i. \end{cases} \quad (17)$$

Here, $r_i$ represents the average distance from the gravity of each frame to the joint $i$ in the training set.

*3.4. Implementation of ST-GCN Based on Label Subset.* The method of ST-GCN in the case of single frame is shown in the formula as follows:

$$f_{\text{out}} = \wedge^{-1/2}(A+I)\wedge^{-1/2} f_{\text{in}} W, \quad (18)$$

$$\wedge^{ii} = \sum_j \left(A^{ij} + I^{ij}\right), \quad (19)$$

where $\wedge^{ii}$ represents the normalized term; A represents the adjacency matrix of the human joint connection; $I$ stands for the self-connected identity matrix; $W$ represents the weight matrix formed by stacking the weight vectors of the output channel.

Considering that there are multiple subsets of labels in practice, the spatiotemporal graph convolution cannot form $\wedge^{-1/2}(A+I)\wedge^{-1/2}$. Therefore, it is necessary that the input performs tensor multiplication with the normalized adjacency matrix, and the result performs the time dimension convolution with the standard convolution of length $1 \times \Gamma$. The input feature graph can be expressed as a $(C, T, V)$ dimensional tensor, where $C$ is $(x, y)$ score, $V$ represents the joint number, and $T$ represents the sequence length. The adjacency matrix can be expressed by multiple matrices $Aj$, namely $(A+I) = \sum_j A_j$. So, formula (20) can be expressed by formula (21), which is shown as follows:

$$f_{\text{out}} = \sum_j \wedge_j^{-1/2} A_j \wedge_j^{-1/2} f_{\text{in}} W_j, \quad (20)$$

$$\wedge_j^{ii} = \sum_k \left(A_j^{ik}\right) + a. \quad (21)$$

To avoid that the denominator is 0, this article sets $a = 0.001$.

## 4. Results and Analysis

*4.1. Experimental Environment and Basketball Movement Classification.* In Python, the results were counted and displayed by using pyqt5 and openCV, and the posture estimation is processed by using OpenPose. Basketball movement classification is the premise of action analysis. This study is based according to the current commonly used basketball Kinetics dataset to classify the basketball movements. Four types of basketball-related actions are obtained, namely running with the ball, layup, pitching, and playing basketball. Among them, playing basketball includes a series of basketball actions, which belongs to multiple basketball action categories. So this experiment only selected three kinds of actions, such as running with the ball, layup, and throwing the ball, as the basketball movement category. In addition, considering the possible state of movement of basketball players on the court, this experiment complements four types of actions: running without the ball, passing the ball, catching the ball, standing, or defending. Finally, the basketball action category in this experiment contains a total of seven kinds of actions, as shown in Table 1.

*4.2. Data Sources and Preprocessing.* In this experiment, video clips of NBA standard games collected by the self-developed basketball action capture gadget are used as the experimental data. Tools include play, stop, fast forward, fast back, and jump to the specified frame function. In addition, there is a tracking algorithm consisting of two parallel forward networks added into the tool, where one network is used to calculate the representation of template features, and the other network is a tracking network. The center point feature and the template feature are used to find the most similar location as the boundary frame.

Considering that the center of the calibration frame of the tracking algorithm is usually target center, and the size of the calibration frame varies with the size of the target. It has a great influence on the target posture extraction. Therefore, this study sets the clipping frame center to the standard frame center and sets its size to 368 × 368 consistent with the network input size. In addition, in order to enlarge the dataset, the captured video is flipped horizontally in this study. At the same time, considering that the calibration tool may have untraceable situations of targets in complex scenes, this paper uses the manual calibration method to track. Finally, the number of videos obtained in this lab is shown in Table 2.

*4.3. Network Structure and Parameter Settings.* The method of basketball motion action analysis based on the spatial temporal graph convolution is constructed in this study. The spatial temporal graph convolution network structure of ST-GCN is designed in Figure 4. In the figure, the left figure is a spatial temporal graph convolution network formed by stacking seven-layer ST-GCN modules. The fourth layer network is used to compress the feature information of the time dimension, and it doubles the number of feature channels. The spatiotemporal dimension convolution step for the convolution kernel of this layer network is 2. The middle figure is a specific form of the ST-GCN module, whose input dimension is $(B, C, T, V, N)$, where $B$ represents the batch size, $C$ represents $(x, y)$ score obtained from the posture estimation model, $T$ represents the sequence length with an initial value of 300, $V = 18$ represents the joint number, and $N$ represents the maximum output number of posture estimation. Since this study only focuses on the central target action, $N$ is set to 1. By multiplying the tensor with its corresponding normalized transformation matrix, it can perform convolution with the general two-dimensional convolution $W_j$.

Furthermore, to achieve basketball movement classification, it is necessary to map the output characteristic information of the ST-GCN module. Here, average pooling is used to compress the output features, and full convolution is used to map the features to seven types of basketball action channels. Finally, the dimensions are changed into (1, 7) for classification.

At last, the experiment sets the temporal dimension graph convolution kernel size of the spatial temporal convolution network to 9. And following the label subset division strategy, the spatial dimension graph convolution kernel size is set as 1, 2, or 3. The initial parameters of the spatial temporal graph convolution network are Kinetics pretraining network parameters of transfer ST-GCN training, and the final classification layer parameters are initialized by the Gaussian distribution. An Adam optimizer is used to update the training process, and the basic learning rate is 0.001. When the 960 epoch is trained, the gradient is decreased by 90% at 320, 480, 640, and 800 epochs.

*4.4. Experimental Results.* To analyze the influence of different frame lengths as input on the recognition effect of the proposed method, and under the premise of other parameters remaining unchanged, the model training is processed with frame lengths of 130, 150, 170, 190, 210, and 230 as input. The results are shown in Figure 5. As can be seen from the table, the recognition effects of most models on motion actions improve with the increase of the frame length, while the recognition effects of some motion actions jump and decline with the increase of the frame length. Overall, the accuracy of Top1 is improved with the increase of the frame length. When the frame length exceeds 190, the recognition effect is not improved because the excessive frame length leads to redundancy. It can be seen that the space is wasted and the effective frame loss is increased. Therefore, this study sets the frame length to 190.

To analyze the influence of label subset division strategy on the motion recognition effect, this paper divided the label subset according to unified division, distance division, and spatial structure division strategy. And the proposed method is adopted for identification. The results are shown in Figure 6. As can be seen from the table, label subsets divided by distance and spatial structure have better effects compared with unified division. The reason is that the subset obtained by unified division is a single subset, which contains less information and has weak information expression ability. However, the subset obtained by distance division and spatial structure division has more information than the subset obtained by unified division, so its effect is better. Compared with the spatial structure division method, the representation by distance division is less, and the action recognition accuracy of the two methods is close.

Table 1: Classification of basketball movement.

| Basketball action category | | | | | | |
|---|---|---|---|---|---|---|
| Pass the ball | Catch a ball | Layup | Pitching | Run without the ball | Run with the ball | Standing or defending |

Table 2: Collection quantity statistics.

| | Pass the ball | Catch a ball | Layup | Pitching | Run without the ball | Run with the ball | Standing or defending |
|---|---|---|---|---|---|---|---|
| Training set | 105 | 91 | 74 | 61 | 145 | 68 | 89 |
| Testing set | 47 | 68 | 18 | 21 | 61 | 19 | 23 |

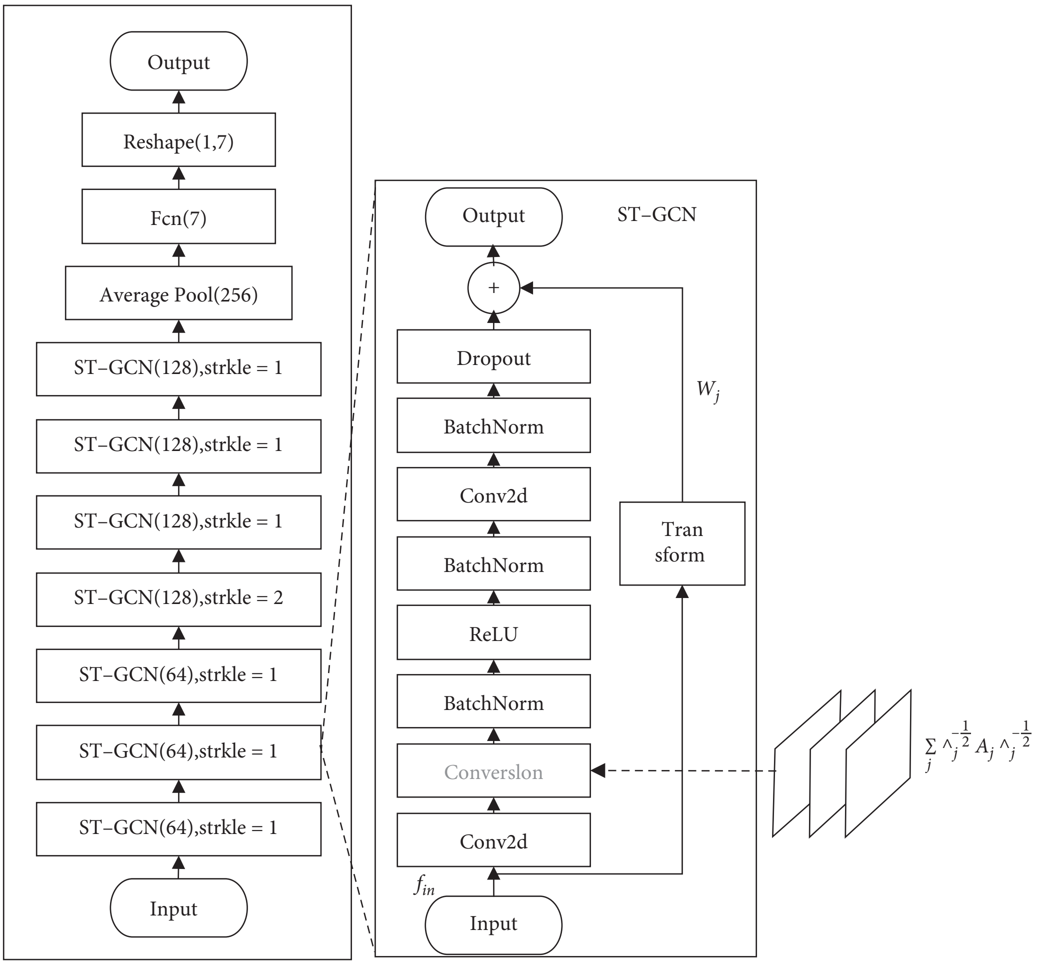


Figure 4: Convolution network structure of migration ST-GCN spatiotemporal diagram.

Therefore, this study chooses the spatial structure division strategy to divide label subsets.

To analyze the influence of different network structures on the model recognition results, different network structures are adopted after the input frame length and label subset division strategy are determined, as shown in Figure 7. Testing the recognition effect of the model on the motion actions, the results are shown in Figure 8. As can be seen from the table, changes in network layers and network structure have a limited effect on improving the accuracy of model recognition results. Compared with the model using the transfer learning method, the accuracy of Top1 is lower. The reason is that the amount of data in the dataset is limited, and more information is not obtained through transfer learning, so its adaptability cannot be effectively improved.

To verify the effectiveness of the proposed method, the proposed method is used to verify it on the experimental dataset. Compared with different motion action recognition methods, the results are shown in Table 3. It can be seen from the table that the method proposed in this study has the best action identification effect in most basketball sports scenes. Although the identification effect of running with the ball is lower than that of the feature descriptor method, the overall

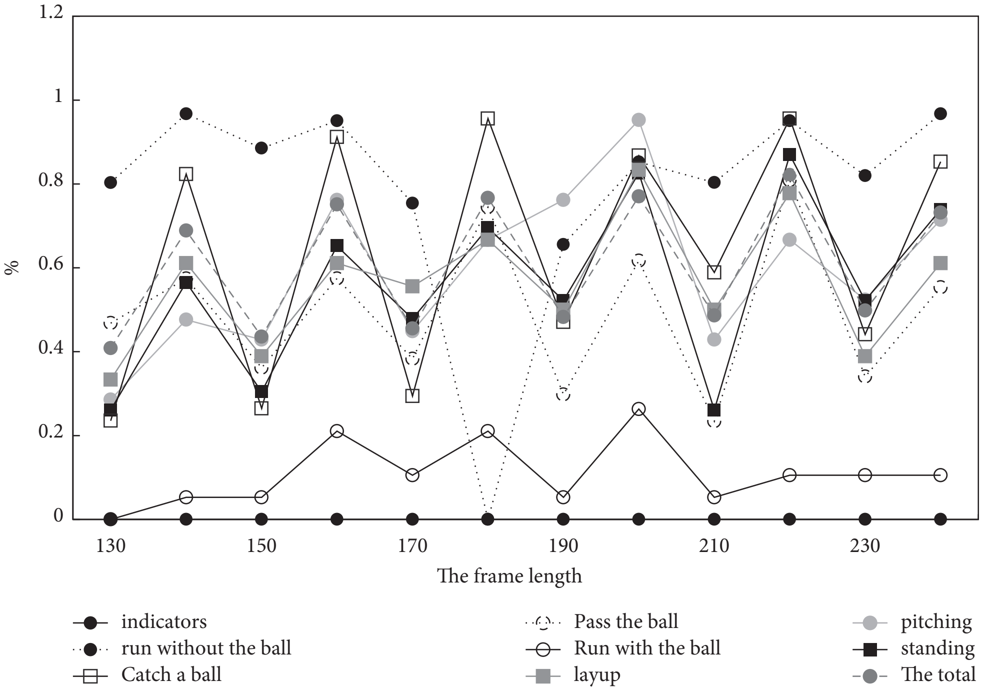


Figure 5: Effects of different frame lengths on recognition results.

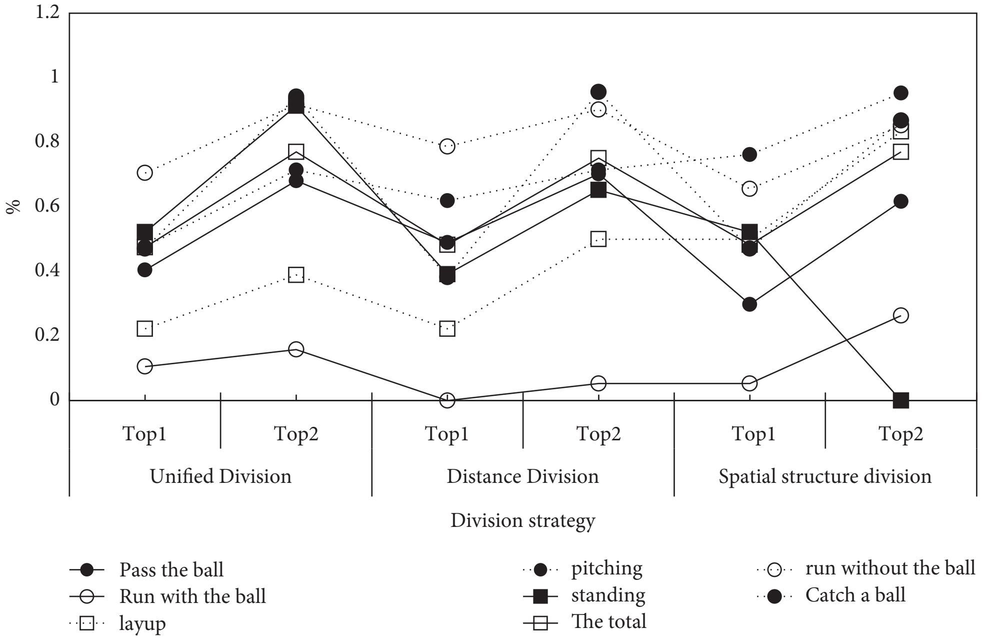


Figure 6: Effects of different partition strategies on action recognition.

action identification effect is better. Therefore, the method proposed in this study is effective to some extent.

Significantly, it can be seen from the test results that the recognition accuracy of two similar movements, running without the ball and running with the ball, is quite different. The recognition accuracy of the proposed method for running without the ball is more than 75%, while that for running with the ball is only about 21%. In order to analyze the causes, this study selects the typical movements of running with and without the ball in the experimental dataset to analyze, as shown in Figure 9. Here, running with the ball and running without the ball are both movements of swinging hands and running with both legs, and the posture joints of the actions are highly similar. Running with the ball has more arm swing than running without the ball. After the images are input into the network and the results of

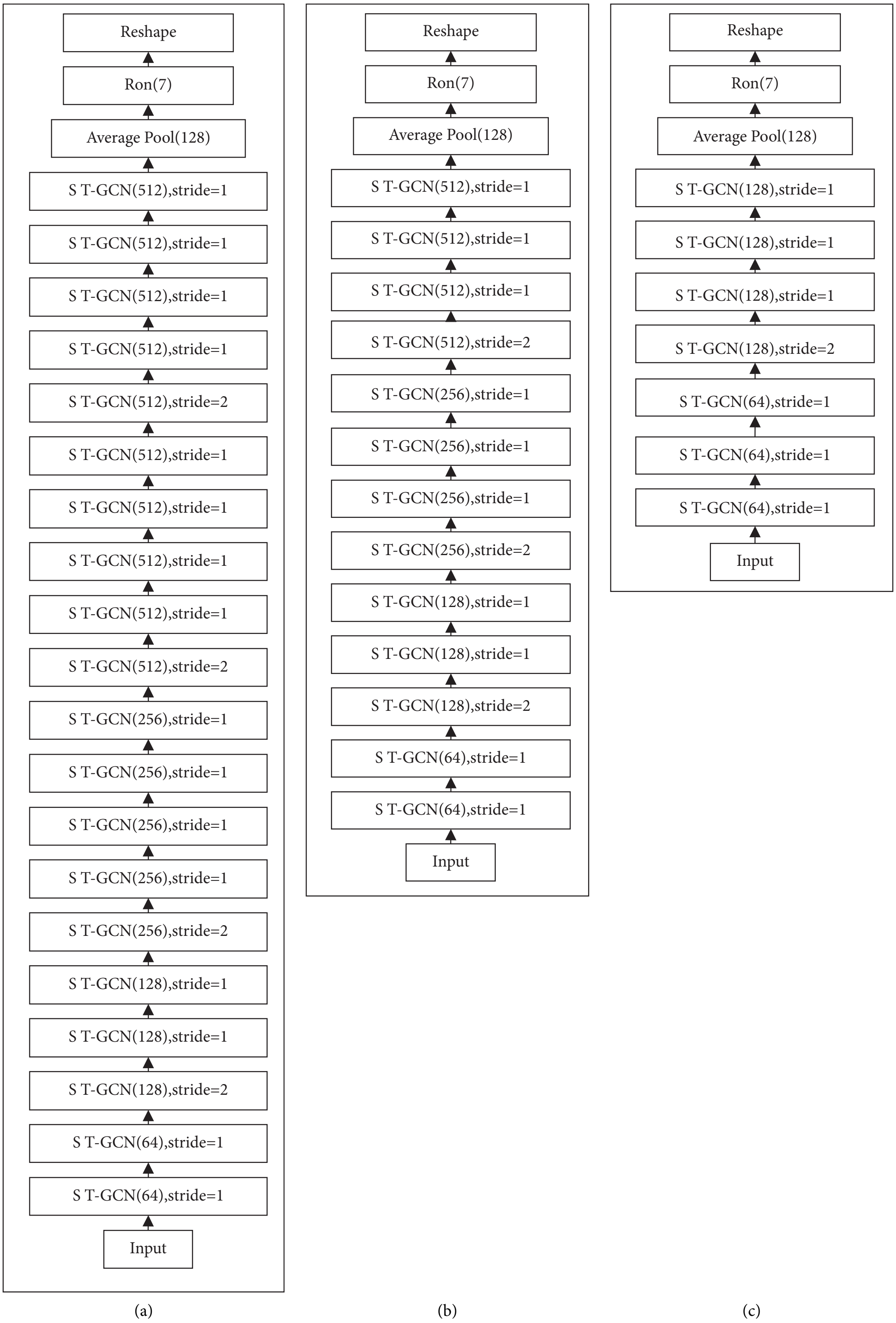


Figure 7: Network structure test. (a) ClassVGG19, (b) 15 storey structure, and (c) 7 storey structure.

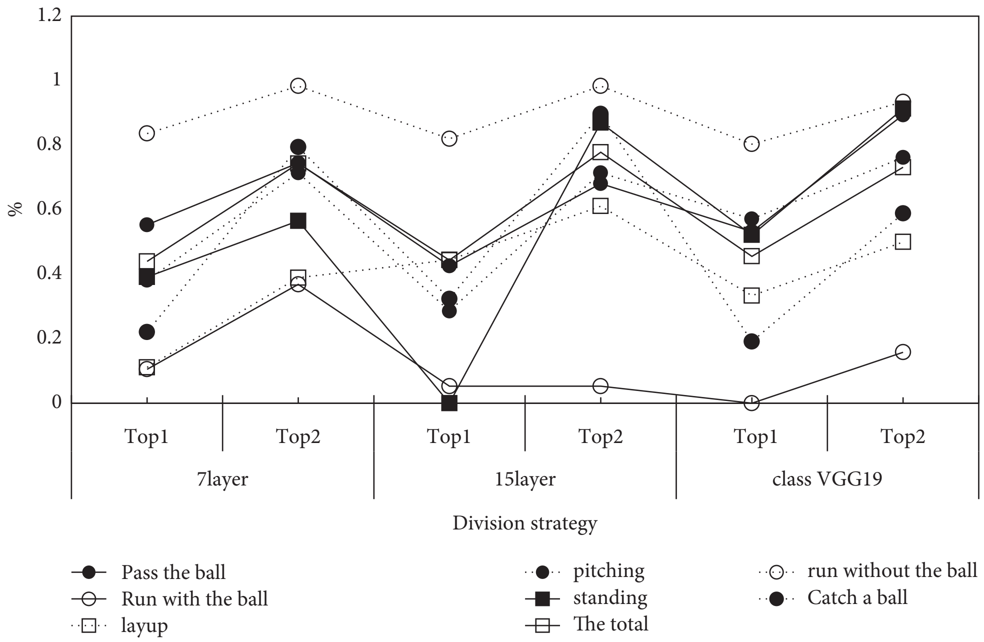
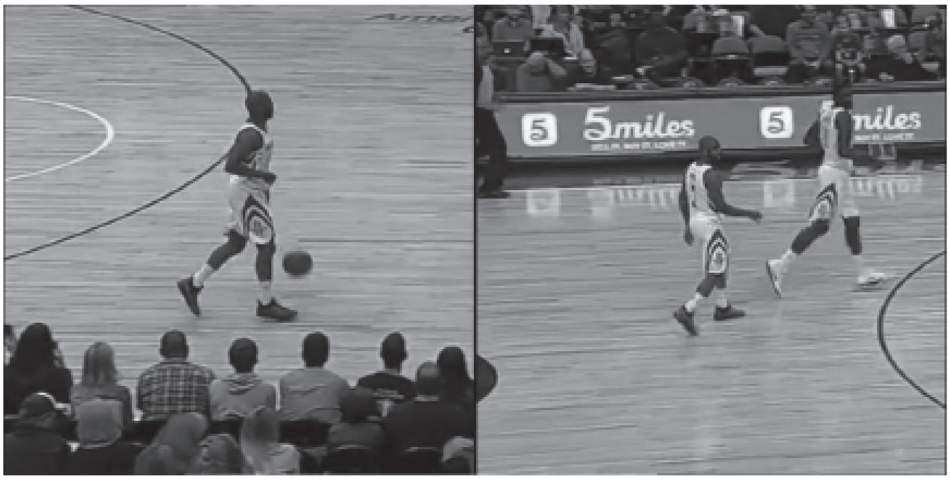


Figure 8: Comparison of test results of network models with different structures.

Table 3: Comparison of identification results of different methods.

| Frame length | Indicators | Pass the ball (%) | Pitching (%) | Run without the ball (%) | Run with the ball (%) | Standing (%) | Catch a ball (%) | Layup (%) | Total (%) |
|---|---|---|---|---|---|---|---|---|---|
| LSTM | Top1 | 21.28 | 23.81 | 47.54 | 5.26 | 17.39 | 14.71 | 22.22 | 24.51 |
| | Top2 | 31.91 | 42.86 | 50.82 | 10.53 | 30.43 | 48.53 | 33.33 | 40.08 |
| Res-CNN | Top1 | 25.53 | 28.57 | 49.18 | 10.53 | 30 43 | 19.12 | 38.89 | 29.96 |
| | Top2 | 51.06 | 47.62 | 555 | 15.79 | 47.83 | 63.24 | 44.44 | 51.75 |
| Paper method | Top1 | 38.30 | 42.86 | 75.41 | 10.53 | 47.83 | 29.41 | 55.56 | 45.53 |
| | Top2 | 74.47 | 66.67 | 83.61 | 21.05 | 69.57 | 95.59 | 66.67 | 76 67 |



(a) (b)

Figure 9: Basketball movement. (a) Running with the ball and (b) running without the ball.

misjudgment are checked, it can be found that the reason for the low recognition failure rate of running without the ball may be that the training data occupy a large proportion in the training set, and the reason for the low recognition failure rate of running with the ball is that it is easy to misjudge it as running without the ball. In addition, the sphere is considered to be added into the posture estimation as a joint. However, for the small amount of calibration data, the recognition effect has not reached the expected standard, so the study has not obtained a satisfactory solution to this problem.

## 5. Conclusion

To sum up, the motion action analysis method at basketball sports scene based on the spatial temporal graph convolutional neural network is proposed. And the human joints and limbs are modeled by using the graph structure in the data structure, and the posture movement is modeled by the spatial temporal graph structure, which realizes the body posture extraction and estimation at the basketball scenarios. The motion fuzzy posture recognition is realized by dividing and applying the tag subset and training with transfer learning. When the spatial temporal graph convolution network has 11 layers, the input length is 190 frames. And when the label subsets are divided by the spatial structure, the network has the highest recognition effect and recognition accuracy in the basketball sports scene, reaching more than 75%.

Compared with other identification methods such as feature descriptors, this method has higher identification accuracy, and it can be used for the motion action identification and analysis in actual basketball sports scenes. Although some achievements have been made in this study, there are still some shortcomings to be improved. Especially, for the low recognition accuracy of running with the ball and easily misjudged as running without the ball, the new identification methods of the ball should be combined to distinguish in the future study so as to improve its recognition accuracy.

## Data Availability

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

## Conflicts of Interest

The author declares no conflicts of interest.

## References

[1] H. Yu, A. Sharma, and P. Sharma, "Adaptive strategy for sports video moving target detection and tracking technology based on mean shift algorithm," *International Journal of System Assurance Engineering and Management*, pp. 1–11, 2021.

[2] L. Liu, G.-h. Chai, and Z. Qu, "Moving target detection based on improved ghost suppression and adaptive visual background extraction," *Journal of Central South University*, vol. 28, no. 3, pp. 747–759, 2021.

[3] W. Li et al., "Moving target detection method based on FPGA," *Scientific Journal of Intelligent Systems Research*, vol. 3, no. 3, 2021.

[4] B. Wu, C. Wang, W. Huang, D. Huang, and H. Peng, "Recognition of student classroom behaviors based on moving target detection," *Traitement du Signal*, vol. 38, no. 1, pp. 215–220, 2021.

[5] L. Huang, "Moving target detection method of three-dimensional image of whip leg technique in s," *Journal of Physics: Conference Series*, vol. 1744, no. 4, Article ID 042216, 2021.

[6] W. Sun, D. Yan, J. Huang, and C. Sun, "Small-scale moving target detection in aerial image by deep inverse reinforcement learning," *Soft Computing*, vol. 24, no. 8, pp. 5897–5908, 2020.

[7] M. Bharat Kumar and P. Rajesh Kumar, "Bayesian fusion strategy for moving target detection in multichannel SAR framework[J]," *Evolutionary Intelligence*, pp. 1–14, 2020.

[8] W. Zhang and W. Sun, "Research on small moving target detection algorithm based on complex scene," *Journal of Physics: Conference Series*, vol. 1738, no. 1, Article ID 012093, 2021.

[9] L. Yaofeng and Y. Ma, "Internet of moving target detection method based on nonparametric background model," *International Journal of Computers and Applications*, vol. 43, no. 2, pp. 193–198, 2021.

[10] Z. Zhao and G. Lu, "Target motion detection algorithm based on dynamic threshold[J]," *Journal of Physics: Conference Series*, vol. 1738, no. 1, Article ID 012085, 2021.

[11] N. A. L. L. A. S. I. V. A. M. Manikandaprabu and S. E. N. N. I. A. P. P. A. N. Vijayachitra, "Moving human target detection and tracking in video frames[J]," *Studies in Informatics and Control*, vol. 30, no. 1, pp. 119–129, 2021.

[12] Q. Yang, W. Shi, J. Chen, and Y. Tang, "Localization of hard joints in human pose estimation based on residual downsampling and attention mechanism[J]," *The Visual Computer*, pp. 1–13, 2021.

[13] W. Chen, Y. Fan, and Ye Zhang, "Dynamic gesture recognition based on iCPM and RNN[J]," *Journal of Physics: Conference Series*, vol. 1684, no. 1, Article ID 012066, 2020.

[14] H. Zhang, H. Dou, and B. Li, "Research on human action recognition algorithm based on sine feature," *Journal of Physics: Conference Series*, vol. 1518, no. 1, Article ID 012024, 2020.

[15] Li-Q. Hu, Z.-Q. Cai, Li-N. Xing, and Xu Tan, "Human action recognition via learning joint points information toward big AI system[J]," *Journal of Visual Communication and Image Representation*, 2019.

[16] M. Khan, Mustaqeem, U. Amin et al., "Human action recognition using attention based LSTM network with dilated CNN features[J]," *Future Generation Computer Systems*, vol. 125, 2021.

[17] J. Chen, R. Samuel, D. Jackson, and P. Parthasarathy, "LSTM with bio inspired algorithm for action recognition in sports videos," *J]. Image and Vision Computing*, Article ID 104214, 2021.

[18] Ye Lei and S. Ye, "Deep learning for skeleton-based action recognition[J]," *Journal of Physics: Conference Series*, no. 1, p. 1883, 2021.

[19] K.-H. Wu and C.-T. Chiu, "Action recognition using multiscale temporal shift module and temporal feature difference extraction based on 2D CNN," *Journal of Software Engineering and Applications*, vol. 14, no. 05, pp. 172–188, 2021.

[20] A. Vijay Anant, K. Deepak, and C. G. Suresh, "Human action recognition using CNN-svm model[J]," *Advances in Science and Technology*, vol. 105, pp. 282–290, 2021.

[21] J. Yang, F. Wang, and J. Yang, "A review of action recognition based on Convolutional Neural Network[J]," *Journal of Physics: Conference Series*, vol. 1827, no. 1, Article ID 012138, 2021.

[22] A. Kumar, S. Kushwaha, and R. Khurana, "Fusing dynamic images and depth motion maps for action recognition in

surveillance systems[J],” *International Journal of Sensors, Wireless Communications & Control*, vol. 11, no. 1, pp. 107–113, 2021.

[23] D. zheng, H. Li, H. Li, and S. Yin, “Action recognition based on the modified t,” *International Journal of Mathematics and Soft Computing*, vol. 6, no. 6, pp. 15–23, 2020.

[24] J. Lee and H. Jung, “TUHAD: taekwondo unit technique human action dataset with key frame-based CNN action recognition[J],” *Sensors*, vol. 20, no. 17, 2020.

[25] S. Hoshino, K. Niimura, and K. Niimura, “Robot vision system for human detection and action recognition,” *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 24, no. 3, pp. 346–356, 2020.