Scholar Publishing Group
International Journal of Sports Technology
https://doi.org/10.38007/IJST.2022.030106
ISSN 2789-7141 Vol. 3, Issue 1: 63-80

Handball Action Evaluation Method and System Based on Deep Learning

Ivay Jacksion*

Xiqin Technology Co., LTD, China

**corresponding author*

Keywords: Deep Learning, Handball Action Evaluation, Handball Action Recognition, Video Data, Convolutional Neural Network

Abstract: Deep learning system aims to use hierarchical model to learn high-level functions from low-level functions. This study mainly discusses the handball action evaluation method and system based on deep learning. In the process of video data acquisition, the binocular RGB camera system is installed on a fixed iron frame 2.2m above the ground and 3M away from the end of the court in order to capture a wide range of field images including the whole plane of the court. In order to complete a variety of handball recognition in complex background environment, this paper builds a convolutional neural network to accept image input and output image categories, and selects Caffe as an open source, efficient and stable deep learning framework. In handball video processing, convolution neural network is used to detect the position of human body in each frame. Then, the pose of the human body in each bounding box is predicted, and the multi approximate pose of each frame is filtered, and a two-dimensional coordinate estimation of human joint is output. Finally, the 2D coordinates of human joints are used as input to fit the 3D coordinates of human joints. Then CNN is used to process the video data and optical flow data at the same time, and the motion evaluation entrance gives the results by calculating the angle at the joint point. In the experiment, the recognition rate is 88.89%. Among them, the recognition accuracy of other movements is 97% except for push and block. This study is helpful to provide positive guidance for handball training.

1. Introduction

In recent years, deep learning has made great progress. However, there is little research on the robustness of learning systems with deep architecture, which needs further research. In particular, mean square error (MSE) is a commonly used optimization cost function in deep learning, which is quite sensitive to outliers (or impulse noise). Robust methods are needed to improve the learning

effect and eliminate the harmful effects caused by outliers, which are ubiquitous in real-world data.

Action evaluation has a wide range of application scenarios and social values. For example, in sports, if athletes want to improve their professional level, they need to know the differences between their actions and standard actions, and carry out targeted training, in this way, athletes usually can not see the whole process of the action, and the feedback information is not comprehensive enough, and the efficiency is not high. Therefore, it is necessary to develop a set of auxiliary training tools to evaluate athletes' action, which will greatly improve the learning efficiency of trainers.

Deep learning applies a hierarchy of hidden variables to nonlinear high-dimensional predictors. Matthew's goal is to develop and train deep learning architectures for spatiotemporal modeling. He uses stochastic gradient descent and pressure drop to perform parameter regularization to train the depth structure, the purpose of which is to minimize the mean square error of the out-of-sample prediction. Although he summarized the direction of future research on deep learning, he did not give specific data in the research [1]. Levine S describes a learning-based hand-eye coordination method for grabbing robots from monocular images. He trained a large convolutional neural network to predict the probability that the task space motion of the gripper will successfully grasp, using only monocular camera images and having nothing to do with camera calibration or current robot pose. Then, he uses the network to servo the gripper in real time to achieve a successful gripping. Although his research method achieves effective real-time control and can successfully capture novel objects, the research process lacks comparative data [2]. He H believes that when the receiver is equipped with a limited number of radio frequency (RF) chains in a beam space millimeter wave large-scale multi-input and multi-output system, channel estimation will be very challenging. In order to solve this problem, he used the approximate message passing (LDAMP) network based on learning noise reduction. In his research, although the receiver is equipped with a small number of RF chains, the LDAMP neural network is also significantly better than the latest algorithm based on compressed sensing [3] . In order to enhance the invariance of deep representations and make them more transferable between various fields, Long M proposed a unified deep adaptation framework for joint learning of transferable representations and classifiers. His research framework includes two interdependent paradigms, namely, unsupervised pre-training for effective training of deep models using deep denoising autoencoders, and supervised fine-tuning for effective use of distinguishing information using deep neural networks. His research although it shows that both learn by embedding the depth representation into the regenerative kernel, the research process is not logical [4].

In the video data collection process, in order to collect images of a large field of view including the entire court plane, the binocular RGB camera system is installed on a fixed iron frame 2.2m from the ground and 3m from one end of the court. In order to complete the recognition of a variety of handballs in a complex background environment, this paper builds a convolutional neural network that accepts image input and output image categories, and chooses the open source, efficient and stable deep learning framework Caffe. In handball video processing, the input video data is first used to detect the position of the human body in each frame of the image using a convolutional neural network. Then, the posture prediction of the human body in each bounding box is performed, and the two-dimensional coordinate estimation of the human joints is output after filtering the posture of each frame of image. Finally, the two-dimensional coordinates of the human joints are used as input to fit the three-dimensional coordinates of the human joints. Then use CNN to process video data and optical flow data at the same time, and the action evaluation entry gives the result by calculating the angle at the joint point.

2. Action Assessment

2.1. Deep Learning

Overall, recent research in deep learning (DL), reinforcement learning (RL) and their combination (deep RL) is expected to revolutionize artificial intelligence [5-6]. The increase in computing power, along with the increase and increase in data storage speed and the decrease in computing cost, has enabled scientists in various fields to apply these technologies to data sets that were previously difficult to process due to their scale and complexity [7]. The degree matrix and the adjacency matrix can be used to calculate the Laplacian matrix. The Laplacian matrix can better reflect the relationship between the nodes. The calculation formula is as follows [8-9]:

$$F = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{1}$$

Among them, D represents the adjacency matrix [10]. Graph convolution is similar to traditional convolution in that the calculation process of graph convolution is also a process of sampling first and then weighting and summing [11-12]. The sampling method of graph convolution is similar to traditional convolution. The traditional convolution samples the center pixel and surrounding pixels and then enters the convolution network, while the graph convolution samples the center node and surrounding nodes and then enters the network [13].

$$F = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} X \omega \tag{2}$$

Among them, F represents the characteristic information contained in each node in the input graph structure [14-15].

2.2. Handball Action Recognition

Handball movement recognition, that is, input a piece of data containing human movements, and let the computer judge the category of the current movement. Action recognition methods can be divided into video data (Video-based) recognition methods and skeleton data (Skeleton-based) recognition methods according to the input object [16]. Video data is the video image data collected by optical sensors (such as cameras). It has good color information and contour information, which can be fully learned through the network model. And video data collection is convenient, and it is relatively easy to construct large-scale data sets [17-18]. However, network models designed based on video data are severely affected by problems such as occlusion, illumination, background, and resolution [19]. Skeletal data is the use of sensor equipment (such as motion capture system, Kinect camera, etc.) to collect the data of the spatial position coordinates of joint points in motion [20-21]. Skeletal data usually contains spatial location information, which has more information than video data, which is conducive to building network models. However, bone data is very dependent on the accuracy of sensor equipment, and data collection is more difficult [22]. In order to facilitate digital processing, the image is first converted from RGB color space to HIS space, and then the space is segmented using the Yuantong distance criterion, and finally the field ratio value is calculated. The field ratio value is given by the following formula:

$$FR(j) = \frac{D_2(j)}{p \times q} \tag{3}$$

The frame rate of change is a physical quantity that describes the speed of the frame movement [23-24]. Therefore, for this feature of video, the frame rate of change is selected as the approximate position of the auxiliary handball. The specific calculation formula of the frame change rate is as follows:

$$FCM(k)=\frac{1}{p\times q\times|o_{\max}|}\sum_{t=1}^{p}\sum_{j=1}^{q}|o(i,j)| \tag{4}$$

Among them, *FCM(k)* represents the frame change rate of the k th image frame [25]. The handball competition venue is greatly affected by the lighting, which makes the main color of the competition venue float within a certain range and constantly change [26].

$$mean=\frac{\sum_{i=i_{\min}}^{i=i_{\max}}hist(i)\times i}{\sum_{i=i_{\min}}^{i=i_{\max}}hist(i)} \tag{5}$$

Among them, *hist(i)* represents color statistics.

$$SE(c)=\exp((1-n_c)/\alpha) \tag{6}$$

Among them, SE is the lens conversion rate, and c is the lens index value. The specific definition of frame motion intensity is as follows:

$$LMI(c)=\frac{1}{n_c\times m_c}\sum_{j=1}^{n_c}m_c(j) \tag{7}$$

Where *LMI(c)* represents the lens motion intensity of the c-th lens, and n represents the total number of frames of the c-th lens.

2.3. Handball Recognition Network Training

When the k-th image is input, the network will output a 2-dimensional vector $O_k=[O_{k1},O_{k2}]$. Then the error between the network output and the label is:

$$L(y_k,O_k)=\log\left(\sum_{j=1}^{2}e^{O_{kj}}\right)-o_{yk} \tag{8}$$

Using this error to reversely train the parameters of the upper layer of the output layer is to obtain the effect of the error on the parameters. Since there is no direct relationship between the error term and the parameter term, it needs to be disassembled through the chain rule:

$$\frac{\partial L}{\partial W_N}=\frac{\partial L}{\partial O_K}\cdot\frac{\partial O_K}{\partial W_N} \tag{9}$$

Suppose the output of the upper layer of the output layer is X_k and the bias of the output layer is b, then:

$$O_k=W_N\cdot X_k+b \tag{10}$$

In the disassembled formula, $\frac{\partial L}{\partial O_K}$ represents the derivative of the Soft max Loss function to O_k, which is calculated as follows:

$$\frac{\partial L}{\partial O_{\mathrm{k}}}=\frac{\partial}{\partial O_{\mathrm{k}}}\left(\log\left(\sum_{j=1}^{2}e^{o_{kj}}\right)-o_{yk}\right)=\frac{e^{o_k}}{\sum_{j=1}^{2}e^{o_{kj}}}-\delta_{ky} \tag{11}$$

Among them:

$$\delta_{\mathrm{ky}}=\begin{cases}1 & k=y\\ 0 & k\neq y\end{cases} \tag{12}$$

Then calculate the derivation of O_{k} to W_N:

$$\frac{\partial O_{\mathrm{k}}}{\partial W_N}=\frac{\partial}{\partial W_N}\left(W_N\cdot X_N+b\right)=X_N \tag{13}$$

Available:

$$\frac{\partial L}{\partial W_N}=\frac{\partial L}{\partial O_K}\cdot\frac{\partial O_K}{\partial W_N}=\left(\frac{\mathrm{e}^{O_K}}{\sum_{j=1}^{2}\mathrm{e}^{O_K}}-\delta_{ky}\right)X_N \tag{14}$$

According to the gradient descent method to update the weight W_N, its change should be along the negative gradient direction of the error. Set the weight change rate to η, and the weight change can be obtained as:

$$\Delta W=-\eta\frac{\partial \mathrm{L}}{\partial W_N}=-\eta\left(\frac{\mathrm{e}^{\mathrm{O_k}}}{\sum_{\mathrm{j}=1}^{2}e^{o_{kj}}}-\delta_{ky}\right)X_N \tag{15}$$

In summary, using the chain rule and gradient descent method, the input weight parameters of the output layer can be updated according to the error value between the network output and the label. Similarly, the steps for updating the weights of other hidden layers of the network according to the error value are the same. The handball recognition network is shown in Figure 1. The network designed in this paper contains 2 convolutional layers, 2 maximum pooling layers, 2 ReLU activation layers, and 2 fully connected layers. The I-th convolutional layer accepts an image as input, and contains 16 convolution kernels with a size of 3×3×3, and the convolution step length is 1 pixel. The feature map (Feature Map) output by the convolutional layer is activated by the first ReLU layer and down-sampling by the first pooling layer, and then input to the second convolutional layer. This layer contains 4 convolution kernels with a size of 3×3×16, and the convolution step length is 1 pixel. The feature map output by the second convolutional layer also passes through the activation layer and the pooling layer, and then is input to the fully connected layer, converted into an output vector of size 2×1, and the result of judging whether the image contains a handball is output .

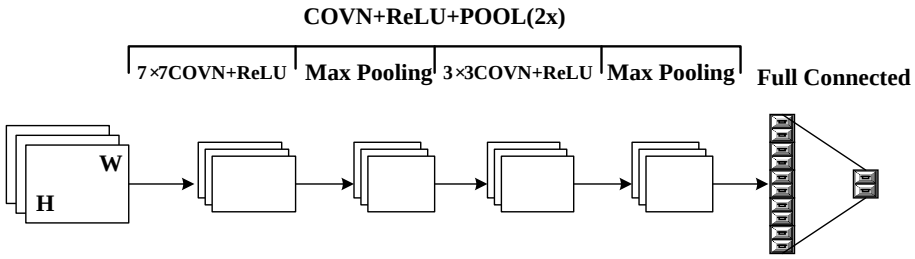

Figure 1. Handball recognition network

3. Handball Action Evaluation Experiment

3.1. Construction of the Video Data Acquisition Platform

The camera vision system uses a binocular industrial RGB camera system for image data acquisition. In order to collect images of a wide field of view including the entire court plane, the binocular RGB camera system is installed on a fixed iron frame 2.2m from the ground and 3m from one end of the court. The frame rate of this model camera can be selected between 120~200fps. In this paper, considering the image transmission speed limit and image resolution requirements, 120fps is selected as the acquisition frame rate to obtain an average of more than 10 images per second.

The self-built handball basic technical action data set is collected by a camera. The camera position is about 4 meters away from the athlete in a straight line, and the vertical distance from the ground is about 2.5 meters and facing the athlete. The 7 types of handball basic techniques that collect 6 players are currently used more frequently in actual combat. When the data set was collected, the athlete approached the table to simulate the action of approaching the table during competitive games. The data set has 4484 video action clips, the resolution of the video is 1280720, and the number of frames is 25 frames per second. This article uses 4035 video action clips as the training set, and the remaining 449 clips as the test set. When the data set II is collected, the athletes are far away from the table to simulate the action of the middle and far stage in the competition. There are 3071 video clips in the data set, of which 2780 video clips are used as the training set, and the remaining 291 clips are used as the test set (data set II). The camera parameters are shown in Table 1.

Table 1. Camera parameters

Model	POINT GREY GRAS- O3K2C-C
Resolution	1920 x1200
Frame rate	1 20-200 fps
Image Sensor	Kodak 0340D CCD
Pixel size	7.4um
Data interface	9 pin IEEE- 1394b
Size	44x29x58mm
Weight	I 04g

3.2. Construction of Handball Recognition Network

In order to complete the recognition of a variety of handballs in a complex background environment, this article builds a convolutional neural network that accepts image input and output image categories, and chooses the open source, efficient and stable deep learning framework Caffe (Convolutional Architecture for Fast Feature Embedding) Build the network on the Internet. Caffe provides a complete toolkit required to build a network, including a modular hierarchical structure, multiple training algorithms, reference models, etc. It also supports CPU and GPU operation, with a streamlined structure and fast running speed. In addition, Caffe also provides C++ Python, Matlab and other language interfaces for joint debugging.

In the network, 1 convolutional layer, 1 activation layer, and 1 pooling layer are connected in order to form a set of typical feature extraction structures. The convolutional layer is responsible for identifying certain features in the graph, and the activation layer is responsible for Non-linear operators are added to increase the richness of features, and the pooling layer is responsible for selecting the features with the largest weight ratio. Since the shape and color features of the handball to be recognized are very prominent, and they maintain high consistency in different images, it can be considered that the feature information that the network needs to extract is relatively concentrated, so the network designed in this paper only contains 2 sets of feature extraction The structure reduces the complexity of the network while not affecting the recognition accuracy.

3.3. Design of Preprocessing Module for Handball Video Data

The first step: Use convolutional neural network to detect the position of the human body in each frame of image for the input video data. Perform multiple convolution operations on the image data and use back propagation to optimize the parameters of the convolution kernel, and continue to stack convolution operations like this to form a neural network-like structure, which is a convolutional neural network. Use a convolutional neural network (such as VGG-16) to output the feature map of the picture (Feature Map). Given the label training parameters of the data, the network will have a higher activation value for the pixels in the area where the human body is located, and the trained feature map is remapped to the original image size through upsampling, and multiple bounding boxes (Bounding Box) can be output, each bounding box represents the possible position of the human body.

Step 2: Predict the pose of the human body in each bounding box, and output a two-dimensional coordinate estimation of the human joints after filtering each frame of image. Simply put, it is to predict the position of the joint points of the characters in the picture, and then connect these joint points, and the convolutional neural network can still achieve this function. Different from the human body detection work that uses the bounding box of the human body position as the label training network, the training label of the human body pose estimation is the position of the human body joint points in the picture. The method of outputting the coordinate positions of human joint points after network training is usually called single-person posture prediction. It is not appropriate to directly input the image in the bounding box of the human position detection output as the pose prediction network. Before inputting the SPPE, the data is preprocessed through a layer of spatial transformation network (Spatial Transformer Network, STN). The working principle of STN is roughly: first generate spatial transformation parameters through a series of network layers (such as fully connected or convolutional neural networks) to realize two-dimensional affine transformation; then resample the input image through this parameter to realize image data The translation, scaling,

or rotation of the network, such processing enhances the robustness of the subsequent network. Input the pictures in the preprocessed bounding box into SPPE, and each bounding box will generate a set of pose estimations of human joint points, and each joint point will have its corresponding confidence. For all the relevant nodes corresponding to each human joint position, only the joint point with the maximum confidence is retained. Such a step is called Non-Maximum-Suppression (NMS). Connect the output joint points to complete the conversion of image data to two-dimensional human body joint point data.

The third step: Use the two-dimensional coordinates of the human joints as input to fit the three-dimensional coordinates of the human joints. By using the coordinates of the human joint points in the three-dimensional space as labels, the two-dimensional coordinates can be converted into three-dimensional coordinates through the network. This is a three-dimensional human pose estimation method for video data based on time convolution on two-dimensional joint point trajectories. This method uses a convolutional neural network based on a residual module to perform temporal convolution on two-dimensional joint points, and constrains the position of human joint points in space, and better fits the three-dimensional coordinates of human joints.

Feature extraction: Build an improved Map Reduce task processor on the distributed deep learning system to process feature extraction of video data. Different from the common data matching, data statistics MapReduce task, the improved Map Reduce task of distributed feature extraction has the following two characteristics: 1) Improved Map Reduce is not based on traditional Hadoop, Spark and other distributed frameworks. Using Map Reduce programming ideas, a new implementation is made on the distributed deep learning system. Distributed feature extraction is implemented by using improved Map Reduce based on general-purpose computing on GPU resources. The feature extraction of each video stage is based on GPU acceleration to do the forward prediction task of deep learning, which is in the same line as the content of the distributed deep learning system constructed in this article. The distributed deep learning system built in this article is also based on GPU clusters. Therefore, the improved MapReduce task can be well run on the distributed deep learning system for model forward feature extraction. 2) Since the subtasks are independent of each other, and no additional converging operations, such as merge sorting, are needed. In the process of large-scale video feature extraction, the feature extraction process of each video stage is independent of each other, and only needs to recover the calculated results without additional operations. Therefore, in response to the need to shield the Reduce process, the Map Reduce used here is a Map-Only computing structure. All feature extraction work is completed only in the Map phase, and all calculations are finally recovered on the server side of the distributed deep learning system. The result is fine.

3.4. Video Data Fusion

A dual-stream fusion method based on video data is adopted. This method first extracts and visualizes the dense optical flow of video data, and uses CNN to process video data and optical flow data at the same time. The feature information encoded using CNN is input into the Soft Max classifier, which classifies each set of data used for testing and outputs the probability that the set of data belongs to a certain category. The output probabilities of the two data streams are averaged to achieve information fusion. The comprehensive recognition rate after information fusion is often higher than that of a single data stream.

3.5. Display of Action Evaluation Results

Through the action evaluation result entry of the main interface, you can view the evaluation result between a test action and a standard action. The action evaluation system developed in this paper uses a dynamic time warping method to align the two sequences for video sequences. Since the overall similarity of the sequence is of little guiding significance to the players, the system still compares and evaluates the differences in posture during the action and displays the results. Generally speaking, when a learner learns an action, the captured action video sequence may be as many as 100 frames. If you compare the difference between your posture and the standard posture frame by frame, it will take a long time and the learner may lose patience, the learning effect is not good. Secondly, it is not efficient to practice the postures one by one. If you break down a wave motion, select several key movements to practice targeted, so that these key movements the posture of "is getting closer and closer to the standard, so the overall waving action will become more and more standard. Compared to correcting frame by frame, this method is more efficient.

Therefore, after the standard movements are collected, the key movements in the waving movements are determined at the same time as a template for the standard movements. The selection criteria and number of key movements can be set more scientifically and reasonably according to the opinions of the handball coach. The selection of key movements is not Uniform standards. This thesis takes the height of the elbow relative to the ground as the selection criterion, and selects four key actions. These four actions can present the entire wave process, corresponding to the start, hit, strike, and casual in the process of handball wave; Every time a test action is collected, the system automatically finds the frames in the test video that are aligned with the key actions in the standard video, without manually marking the locations of the key actions one by one. During training, learners can view the difference between their own posture and the standard posture through posture analysis. The joint angle can reflect to a large extent whether an action is close to the standard action. Therefore, the evaluation of this system is by calculating the joint points,angle to give the result.

4. Handball Action Evaluation Results

4.1. Handball Recognition Analysis

The error and accuracy changes during training are shown in Figure 2. Since images with smaller scales are faster during training, this article first trains the network with the collected image data set of 64×48. In this data set, 5000 images are randomly selected as the training data set (Train Set), and another 1000 images are randomly selected as the test data set (Test Set). After training on the training data set, the network converges after 6000 iterations, and the training error is less than 0.2%. In the training process, a test is performed after every 100 iterations, and the test accuracy reaches more than 99%. After training the network with the small image data set, without changing the network structure, directly use the initially trained parameters as the initial value of the network, and further use the large image data set to fine-tune the network to speed up the large image data set Training speed. In this data set, this article also randomly selected 5000 images as the training data set, and the other 1000 as the test data set. The network can converge after 2000 iterations, the training error is less than 0.2%, and the test accuracy is more than 99%.

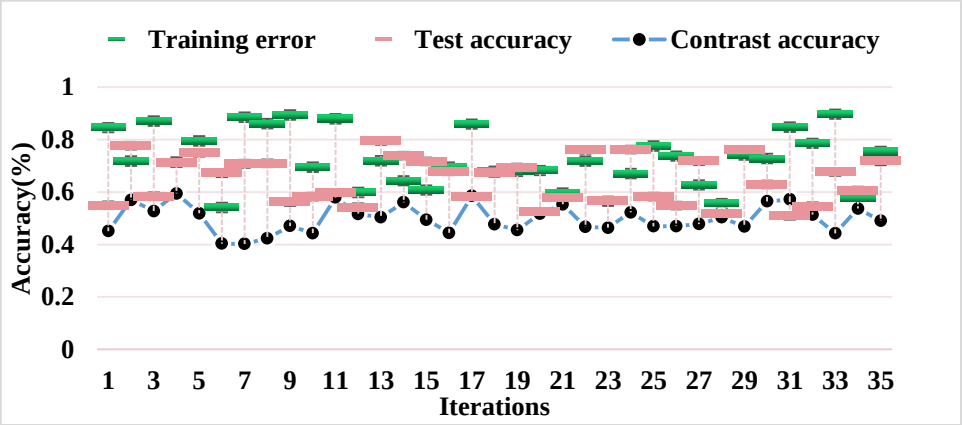

Figure 2. Error and accuracy changes during training

Table 2 shows the comparison between accuracy and recognition time under different network structures. The comparison results of accuracy and recognition time under different network structures are shown in Figure 3. After preliminary training, the handball recognition network proposed in this article has fully met the recognition accuracy requirements. In order to make the network reach the standards of high precision and high efficiency at the same time, this paper designs a set of comparative experiments to find a network structure that can complete the identification faster without sacrificing accuracy by comparing the accuracy and running time of the network under different structures. . A total of 6 network structures are designed in the experiment, and the level collocation is shown in Table 3.1. Each kind of network has passed 6000 trainings of 5000 large image data sets, without any pre-training measures, and then used 1000 large images as tests to calculate the average recognition accuracy and average recognition time of these 100 tests. Each training and test runs on the same x86 host equipped with TeslaK40cGPU. It can be seen from Figure 3 that with the decrease in the number of convolutional layers (such as the comparison between the first group and the sixth group) and the decrease in the number of fully connected layer nodes (such as the comparison between the first group and the third group), the recognition time of the network is significant Decrease, but the recognition accuracy also decreases. In the 6 groups of experiments, the third group of networks achieved a recognition time of 1.55ms and maintained accuracy above 99% by appropriately reducing the number of nodes in the fully connected layer and maintaining a two-layer convolutional layer structure. Considering comprehensively, this article believes that the third group of network structure is the optimal structure of the current network.

Table 2. Comparison of accuracy and recognition time under different network structures

Serial number	Conv1[1] size	Conv2[1]	Conv2[1] core size	fc1[1]number of nodes	Accuracy	Operation hours
1	3x3	8	3x3	96	99.5%	1.87ms
2	5x5	8	3x3	96	98.1%	1.89ms
3	3x3	8	3x3	32	99.19%	1.55ms
4	3x3	8	3x3	16	95.6%	1.40ms
5	3x3	-	-	32	90.3%	0.94ms
6	3x3	-	-	96	90.5%	0.97ms

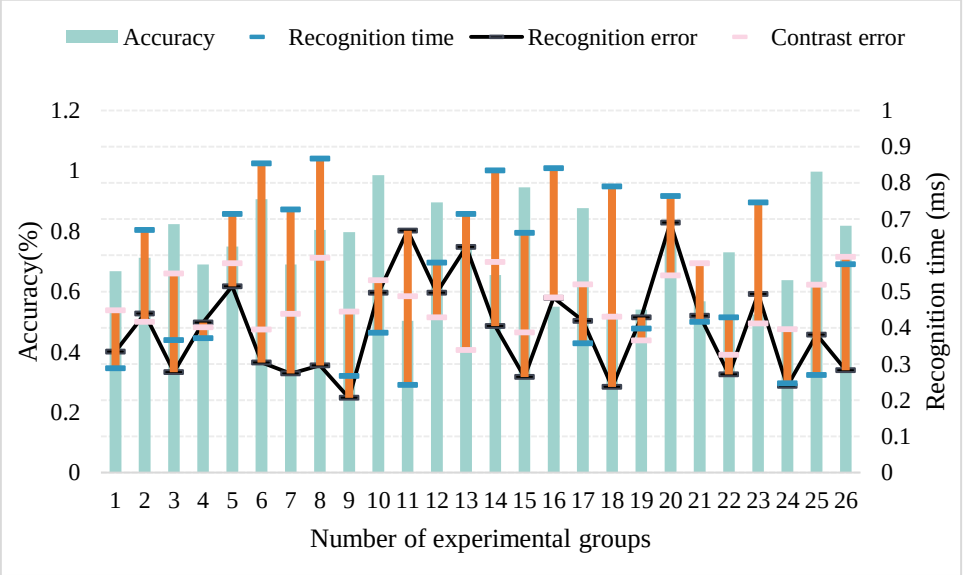

Figure 3. Comparison results of accuracy and recognition time under different network structures

The test data set used in the research also comes from the large image data set collected, and 500 images are randomly selected from the images of category 1 (that is, containing handball) to form a positive example set, and the images of category 0 (that is, without handball) are selected. 500 pieces are randomly selected to form a counterexample set. The pros and cons include three handballs and five lighting conditions. Analyzing and comparing experimental data, it can be seen that the recognition speed of the color-based contrast method is faster, but because the data set contains handball images under multiple colors and multiple lighting conditions, this method has poor adaptability and can only accurately identify a single color and fixed The ball in the environment, so the recall rate of recognition is low. The neural network-based recognition method proposed in this paper can accurately recognize handballs in a data set with multiple colors and multiple illuminations, with high recall and precision. However, due to the complicated operating mechanism of the neural network, the recognition speed of this method is lower than that of the comparison method. But considering that the frame rate of the visual system is 120Hz, the visual perception algorithm of handball only needs to complete all operations within s=8.3ms, so the recognition speed of this method is sufficient to meet the real-time requirements. After testing a total of 1000 pictures, the average recognition results of the two methods are shown in Table 3, and the average recognition speed is shown in Table 4. Classification and recognition based on color features are shown in Figure 4.

Table 3. The average recognition results of the two methods after testing a total of 1000 images

Real category	Identify category			
	Ball		No ball	
	This method	Comparison method	This method	Comparison method
Ball	99.2%	46.2%	0.8%	53.8%
No ball	0.6%	9.2%	99.4%	90.8%

Table 4. Average recognition speed

Parameter	Network-based approach	Color-based approach
Average recognition time/ms	1.55	1.02

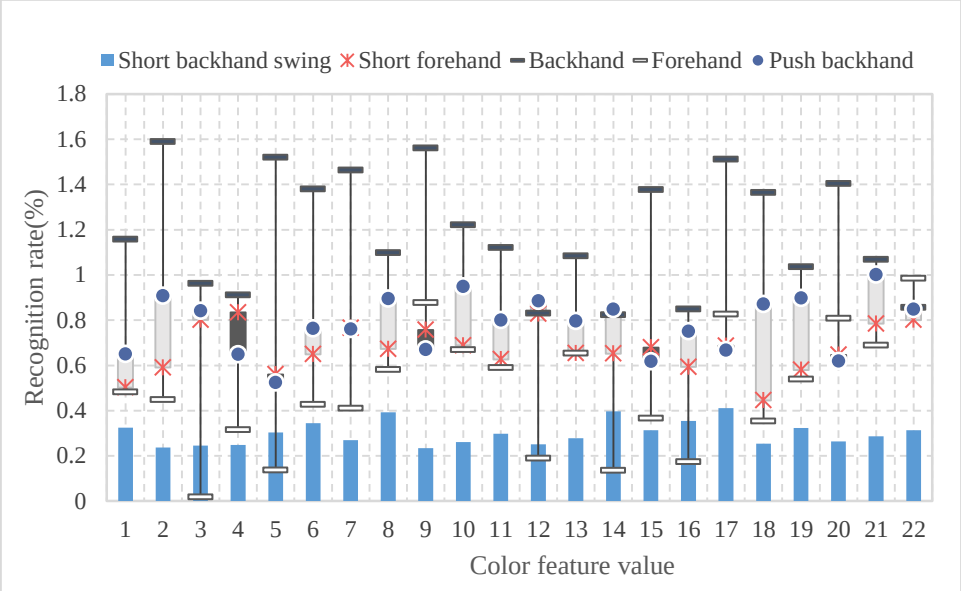

Figure 4. Classification and recognition based on color features

4.2. Methods Combining Video Data and Bone Data

In order to verify the effectiveness of the handball basic technical action recognition method in this paper, the training set of the handball technical action data set is used to train the network model in the text, and the test set is used to test the accuracy of the model. Research has proved that the recognition rate of using cropped video as input is 88.89%. Among them, the recognition accuracy of other actions except for the block push has reached 97%. But the block push is easy to confuse with other actions. The reason may be that the block push is an action that pushes the ball forward after slightly pressing the ball. When the video capture device is in front of the player, the arm forward movement is less obvious than other movements, and it is more susceptible to interference from light and other factors. In order to verify this hypothesis, this paper visualizes the attention map of some test data. From the research results, it can be seen that the attention heat maps of other actions are mainly focused on the hand, and the block action can only be paid attention to when the arm is extended forward. In the horizontal comparison of different algorithms, the methods of TSN (recognition rate 60.32%) and Attention is AII We Need (recognition rate 66.97%) both use the entire video image data as input, and the uncropped video image contains more complex backgrounds. Information, so the recognition accuracy is low. The accuracy of the two is shown in Figure 5. The classification accuracy of each type of action using only video image data is shown in Figure 6.

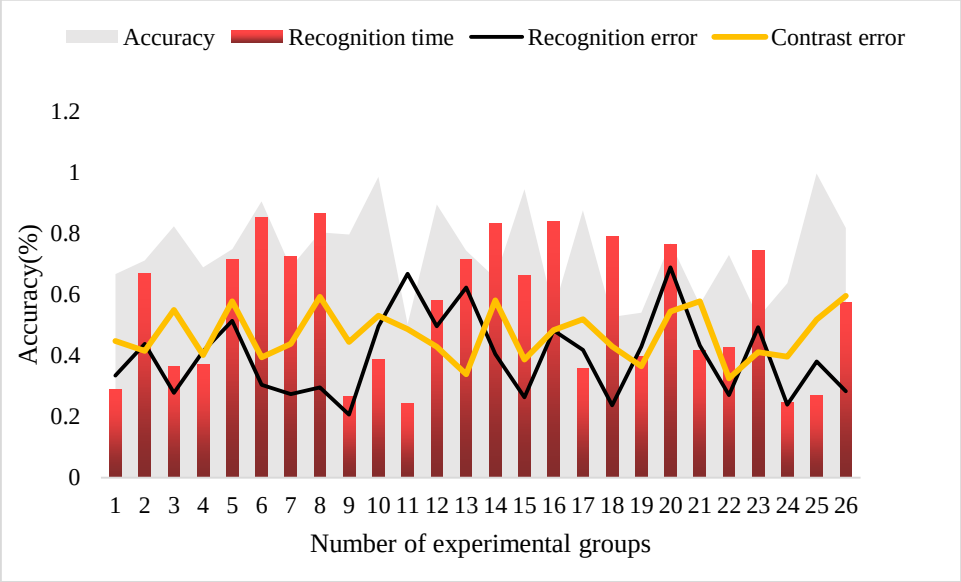

Figure 5. The accuracy of both

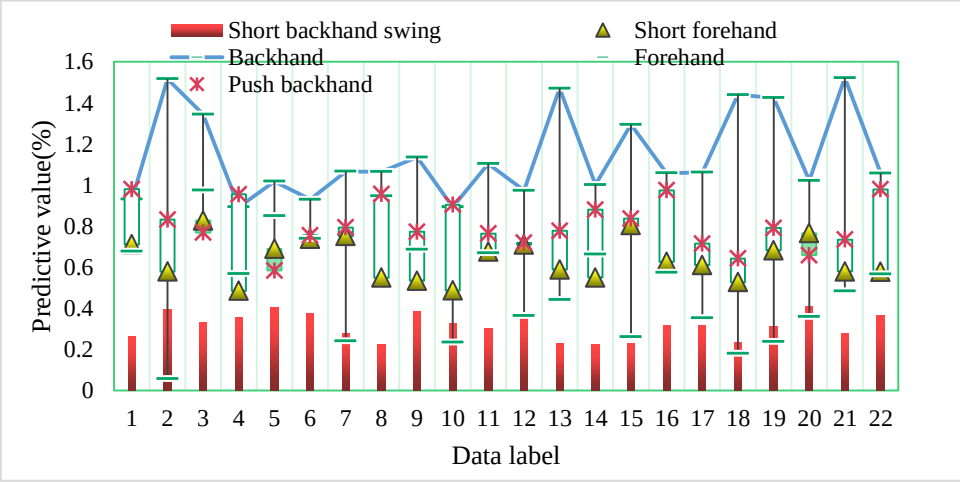

Figure 6. Classification accuracy of each type of action using only video image data

Using self-built basic handball movement data set as shown in Table 5, the method in this paper compares the accuracy of basic handball movement classification with other methods. The recognition rate of the dual-stream fusion method combining video data and bone data reaches 97.97%, which is improved compared to the recognition rate using only video data (88.89%) and the recognition rate using only bone data (93.76%). The reason may be that for some actions, the recognition method based on video data is not easy to distinguish, while the recognition method based on bone data can be better classified. Both of them take advantage of each other when the probability is average, which improves the final recognition rate. Figure 7 shows the comparison between the accuracy of the action classification of this method and other similar methods.

Table 5. Use self-built handball basic action data set

Method	Video data	Bone data	Recognition rate
TSN	Yes	no	60.32%
Attention is All We Needl	Yes	no	66.97%
ST-GCN	no	Yes	70.32%
Recognition video data	Yes	no	88.89%
Recognition local joint points	no	Yes	93.76%
Dual stream integration	Yes	Yes	94.97%

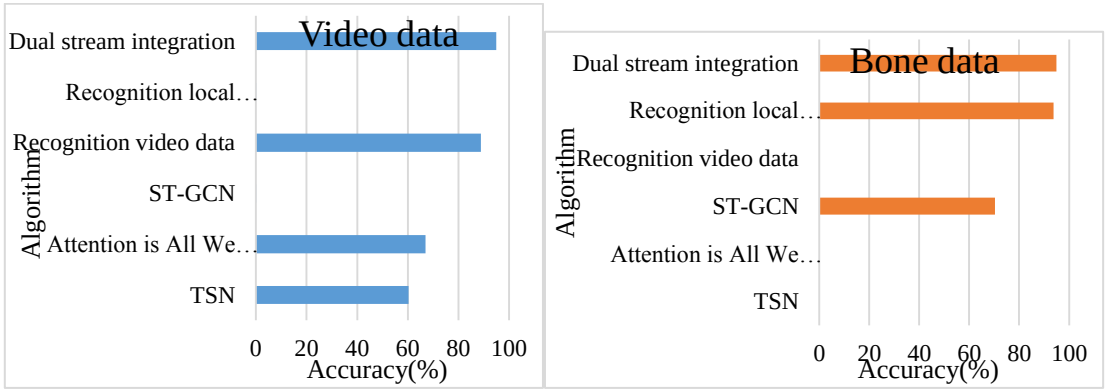

Figure 7. Comparison of the accuracy of the action classification of this method and other similar methods

4.3. Video Action Recognition

The comparison results of all handball video action recognition models based on the original RGB images on the dataset UCF-101 and the dataset HDMB-51. The two most important comparison objects are the LRCN method and the C3D method that are highlighted above. In the model based on the original RGB image of the data set UCF-101, the iRCN model can obtain the highest correct rate, which is 13.9% higher than that of the traditional LRCN model, and 3.3% higher than that of the C3D model. At the same time, on the data set HDMB-51, the accuracy of the iRCN model based on the original RGB image can be improved by 5.2% than the C3D model, while the accuracy of the LRCN model on this data set has not been clearly announced. Table 6 shows the correct rate comparison between UCF-101 and HDMB-51. The UCF-101 recognition result is shown in Figure 8. The HDMB-51 recognition result is shown in Figure 9.

Table 6. Correct rate on UCF-101 and HDMB-51

Method	UCF-101 (%)	HDMB-51 (%)
CNN(mageNet)+SVM	68.8	40.5
LRCN(RGB)	71.7	-
LSTM composite modelQRGB)	75.8	44.0
C3D+SVM	82.3	51.5
3D_ CNN ,STM	83.9	55.2
iRCN(3D_ CNN+biLSTM)	85.6	56.7
LRCN(optical flow)	82.9	-

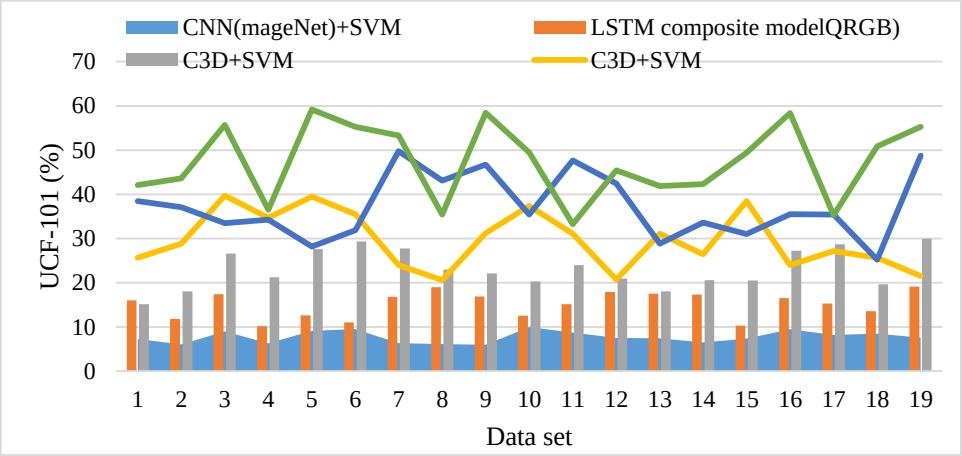

Figure 8. UCF-101 recognition results

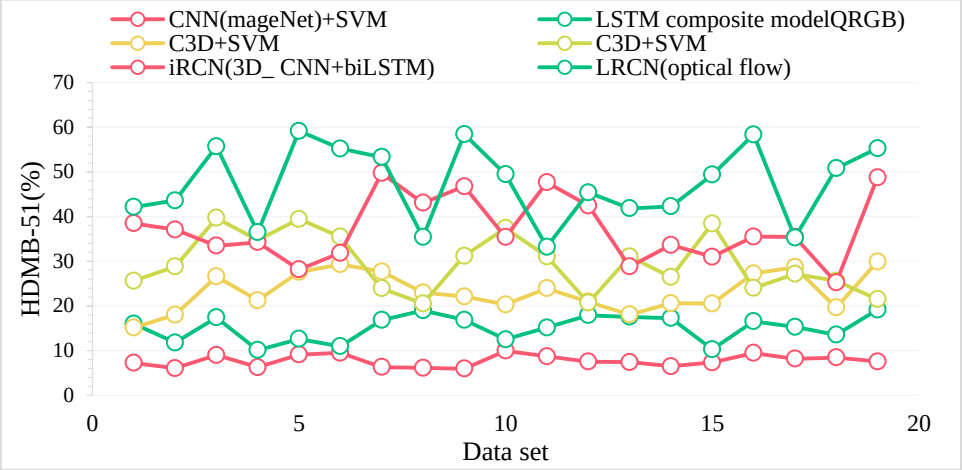

Figure 9. HDMB-51 recognition result

5. Conclusion

For the input video data, a convolutional neural network is used to detect the position of the human body in each frame of the image. Perform multiple convolution operations on the image data and use back propagation to optimize the parameters of the convolution kernel, and continue to stack convolution operations like this to form a neural network-like structure, which is a convolutional neural network. Use a convolutional neural network (such as VGG-16) to output the feature map of the picture (Feature Map). Given the label training parameters of the data, the network will have a higher activation value for the pixels in the area where the human body is located, and the trained feature map is remapped to the original image size through upsampling, and multiple bounding boxes (Bounding Box) can be output, each bounding box represents the possible position of the human body.

Predict the posture of the human body in each bounding box, and output a two-dimensional coordinate estimation of the human joint after filtering the posture of each frame. Simply put, it is to

predict the position of the joint points of the characters in the picture, and then connect these joint points, and the convolutional neural network can still achieve this function. Different from the human body detection work that uses the bounding box of the human body position as the label training network, the training label of the human body pose estimation is the position of the human body joint points in the picture. The method of outputting the coordinate positions of human joint points after network training is usually called single-person posture prediction.

In this study, the height of the elbow relative to the ground was selected as the selection standard, and four key actions were selected. These four actions can present the entire wave process, corresponding to the start, hit, drop, and casual during the handball wave; Every time a test action is collected, the system automatically finds the frames in the test video that are aligned with the key actions in the standard video, without manually marking the locations of the key actions one by one. During training, learners can view the difference between their own posture and the standard posture through posture analysis. The joint angle can reflect to a large extent whether an action is close to the standard action. Therefore, the evaluation of this system is by calculating the joint points,angle to give the result.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Matthew, F, Dixon, et al. Deep learning for spatio‐temporal modeling: Dynamic traffic flows and high frequency trading. Applied Stochastic Models in Business & Industry, 2019, 35(3):788-807. https://doi.org/10.1002/asmb.2399

[2] Levine S , Pastor P , Krizhevsky A , et al. Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection. International Journal of Robotics Research, 2016, 37(4-5):421-436. https://doi.org/10.1177/0278364917710318

[3] He H , Wen C K , Jin S , et al. Deep Learning-based Channel Estimation for Beamspace mmWave Massive MIMO Systems. IEEE Wireless Communications Letters, 2018, 7(5):852-855.

[4] Long M , Wang J , Cao Y , et al. Deep Learning of Transferable Representation for Scalable Domain Adaptation. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(8):2027-2040.

[5] Rojas-Barahona, Maria L . Deep learning for sentiment analysis. Language & Linguistics Compass, 2016, 10(12):205-212. https://doi.org/10.1111/lnc3.12228

[6] Chen L , Qu H , Zhao J , et al. Efficient and robust deep learning with Correntropy-induced loss function. Neural Computing and Applications, 2016, 27(4):1019-1031.

https://doi.org/10.1007/s00521-015-1916-x

[7] Niko Sünderhauf, Brock O , Scheirer W , et al. The Limits and Potentials of Deep Learning for Robotics. The International journal of robotics research, 2018, 37(4-5):405-420. https://doi.org/10.1177/0278364918770733

[8] Maulana G M , Arruda S D M . Teacher assessment action in mathematics classes: a study with teachers from the 2nd cycle of Mozambican general high school education. Acta Scientiae, 2020, 22(5):102-121.

[9] Dokhanchi K . Aftermath of the Gulf War: An Assessment of UN Action, by Johnstone Ian. (International Peace Academy, Occasional Paper Series) 84 pages, notes. Boulder & London: Lynne Rienner Publishers, 1994. $7.95 (Paper) ISBN 1-55587-487-8 Crisis in the Arabian Gulf: An Independent Iraqi View, by Ali Omar. 168 pages, bibliography, index. Westport, Connecticut & London: Praeger Publishers, 1993. $49.95 (Cloth) ISBN 0-275-94158. Archives of Microbiology, 2016, 144(1):102-104.

[10] Magee S R , Eidson-Ton W S , Leeman L , et al. Family Medicine Maternity Care Call to Action: Moving Toward National Standards for Training and Competency Assessment. Family medicine, 2017, 49(3):210-217.

[11] Islam M T , Ayatollahi S A , Zihad S M N K , et al. Phytol anti-inflammatory activity: Pre-clinical assessment and possible mechanism of action elucidation. Cellular and Molecular Biology, 2020, 66(4):264-269. https://doi.org/10.14715/cmb/2020.66.4.31

[12] Kumar S , Sahu D , Mehto A , et al. Health Inequalities in Under-Five Mortality: An Assessment of Empowered Action Group (EAG) States of India. Journal of Health Economics and Outcomes Research, 2020, 7(2):189-196. https://doi.org/10.36469/jheor.2020.18224

*[13] Muzhikov V , Vershinina E , Muzhikov R , et al. The Method Of Individual Assessment Of The Action Of Insulin And Its Adequate Dose In Diabetes Mellitus *Corresponding Author. World Journal of Pharmaceutical Research, 2019, 8(7):176-205.*

[14] Jekabsone A , Kamenders A , Rosa M , et al. Assessment of the Implementation of Sustainable Energy Action Plans at Local Level. Case Study of Latvia. Environmental and Climate Technologies, 2019, 23(2):36-46. https://doi.org/10.2478/rtuect-2019-0053

[15] K, Alysse, Bailey, et al. Applying Action Research in a Mixed Methods Positive Body Image Program Assessment With Older Adults and People With Physical Disability and Chronic Illness:. Journal of Mixed Methods Research, 2019, 14(2):248-267.

[16] Maoxin W . Suboptimal Interobserver Agreement Among Cytopathologists in Assessment of Pancreatic Lesions: A Call for Action. Clinical Gastroenterology & Hepatology, 2018, 16(7):1040-1042. https://doi.org/10.1016/j.cgh.2017.12.043

[17] Maoxin W . Suboptimal Interobserver Agreement Among Cytopathologists in Assessment of Pancreatic Lesions: A Call for Action. Clinical Gastroenterology & Hepatology, 2018, 16(7):1040-1042. https://doi.org/10.1016/j.cgh.2017.12.043

[18] Tsang A , Wong K , Ryan D , et al. Using an Evidence-Informed Framework and a Self-Assessment Tool to Drive Priority Setting and Action toward Senior-Friendly Care. Healthcare quarterly (Toronto, Ont.), 2018, 21(1):25-30.

[19] Ismaidar, Sumarno, Dwintoro, et al. Legal assessment on criminal sanctions in criminal action of corruption based on justice values. International Journal of Civil Engineering and Technology, 2018, 9(11):680-692.

[20] Tofan N , Andrian S , Stoleriu S , et al. The Assessment of the Surface Status Following the Action of Some Acidic Beverages on Indirect Restorative Materials. Materiale Plastice, 2018, 55(1):129-135. https://doi.org/10.37358/MP.18.1.4978

[21] Thomas K H , Mcdaniel J T , Haring E L , et al. Mental health needs of military and veteran women: An assessment conducted by the Service Women's Action Network.. Traumatology, 2018, 24(2):104-112. https://doi.org/10.1037/trm0000132

[22] Mohammed A , Othman M F , Omar R . Ecowas, Good Governance and Collective Military Action in Liberia: A Post Conflict Assessment. International Journal of Humanities and Social Science, 2017, 22(7):82-90.

[23] Ayyoub A A . An Action Research Approach for Using Self/Peer Assessment to Enhance Learning and Teaching Outcomes. Journal of Teaching and Teacher Education, 2017, 5(1):33-42. https://doi.org/10.12785/jtte/050104

[24] Kourosh A , Hasan S , Parsi S , et al. P013 Anaphylaxis recognition and action: needs assessment-quality improvement empowers tertiary care allergy clinic team. Annals of Allergy Asthma & Immunology, 2016, 117(5):S26-S26. https://doi.org/10.1016/j.anai.2016.09.021

[25] Amit R , Pushpa P . Assessment of Mechanism of Action of Antidiabetic Activity o f Calocybe indica by Enzyme Inhibitory Activity. Biosciences Biotechnology Research Asia, 2016, 13(4):2117-2123. https://doi.org/10.13005/bbra/2372

[26] A J C M , B J F , C H E H , et al. An Ecological Assessment of the Northern Yellowstone Range: Synthesis and Call to Action. Rangelands, 2018, 40(6):224-227. https://doi.org/10.1016/j.rala.2018.10.007