Intelligent Computing Technology and Automation
Z. Hou (Ed.)

doi:10.3233/ATDE231218

Sports Teaching Action Recognition Based on Hybrid CNN-HMM

Xinying XIE[a,1], Dongfeng DING[b]
[a] *Department of Physical Education and Research, Xinjiang University, Xinjiang, 830046, China*
[b] *School of Information Science and Engineering (School of Cyberspace Security), Xinjiang University, Urumqi, 830046, China*

Abstract. In order to automatically recognize abnormal actions in physical education teaching in intelligent monitoring system, a CNN and HMM based action recognition model is proposed. In this scheme, the camera and computer are used to capture and measure the target, and the preliminary features of the image acquired from the data are extracted. Then, hybrid CNN-HMM is introduced as the key technology and main method of behavior recognition, and principal component analysis is also used to reduce the dimension of the extracted feature parameters. Finally, the test set is input into the trained classifier for action recognition, which provides an auxiliary reference for teaching actions. The experimental results show that the system can accurately quantify the human movements in sports teaching, and the recognition accuracy of complex databases with different scenes is better than that of similar methods.

Keywords. sports teaching; CNN; HMM; action recognition; feature extraction

1. Introduction

Since action processing is the key technology of sports, to standardize athletes' actions and make their sports achievements reach or close to the highest level, scientific methods are needed. In order to achieve this goal, with the support of relevant departments, the standard movements of various sports such as basketball, volleyball, track and field have been established, a multimedia sports standard movement database has been established, and fuzzy mathematics and neural network technology have been applied to graphic processing and pattern recognition to achieve computer quantitative analysis and pattern recognition of sports movements [1]. Sports action recognition means the process of tracking some key points in the time domain to record human movement, and transforming it into a usable mathematical way to express the movement. It is of great significance for competitive training and national fitness. Traditional motion recognition technologies include mechanical, acoustic, electromagnetic, optical, etc. Mechanical technology uses external sensors and rigid supports, which will affect limb movement. Acoustic and electromagnetic technologies are vulnerable to external environmental interference, have large time delay, and have low test accuracy. The traditional optical technology is more accurate, but there are also some shortcomings, such as high price, long data processing time, etc. [2]. At present, the technology is still in the research and development stage, and there are still many

[1] Corresponding Author: Xinying Xie; Department of Physical Education and Research, Xinjiang University, Xinjiang, 830046, China; fengjianqiangsc@163.com

problems in the application process. However, it is helpful for students to exchange the input of learning information and simulation scenes.

This paper studies the image and video decomposition technology for physical education teaching. Firstly, the technical platform for motion recognition under sports vision is analyzed, and the principles of image processing and target detection are discussed. Combined with the actual demand of action recognition, it is proposed to use hardware analog sensors to collect data and send it to Flash for further processing. In action recognition, a hybrid CNN-HMM algorithm is applied to decompose action images and extract features for matching through similarity detection to obtain the final recognition result. The recognition method is divided into two stages: the training stage and the testing stage, to acquire an action recognition algorithm model with generalization ability. The simulation experiment takes the teaching demonstration of gymnasts as an example to verify the effectiveness of this scheme through the decomposition and recognition of actions. The results show that the method has strong anti-interference ability and good classification accuracy.

2. Motion Recognition Technology in Sports Vision

2.1 Analysis of Current Situation of Sports Vision Based Motion Recognition System

Because of the relationship between motion recognition and body sensing technology, recognition system is an important part of body sensing technology. It mainly refers to the recording of the movements of the athletes and the interaction of the relevant equipment when the machine is under the control of no one. The recognition of the motion recognition system is mainly based on the angle and characteristics between the bones and joints, which can not only meet the extension in the recognition process, but also can match the movements with high accuracy. Expand the recognition of new actions and improve the algorithm of motion recognition. Motion recognition is to match and recognize the continuous motion of the image under vision, including the temporal order and spatial information of the image. Then, action recognition is to distinguish the phase actions of the image, and the image does not have motion behavior over time. In order to accurately apply the somatosensory technology in real life, it is necessary to improve the accuracy of motion recognition algorithms and sensors [3]. Some scholars made scientific guidance according to people's own needs for exercise and fitness, studied body sensing technology and motion recognition algorithms, built a sports system, collected sports data for analysis, developed scientific sports methods, and improved the server and customer service.

2.2 Image Preprocessing and Moving Object Detection

The main purpose of moving object detection is to extract moving objects from video images and obtain their feature information, such as color, shape, contour, etc. The process of extracting moving objects is actually a process of image segmentation [4,5]. Every joint in the human body has its own potential for movement. For example, the shoulder joint has a huge degree of freedom and range of motion, and the shape features of the detected motion region are also very complex. The program is used to convert the image sequence into video of any resolution, extract the simple moving background image, and then standardize the extracted moving human body image. On

the same data set configuration, connectivity analysis and edge extraction of different sizes are introduced for comparison, and the final image preprocessing and moving target detection data sets have been obtained. The specific process is shown in figure 1.

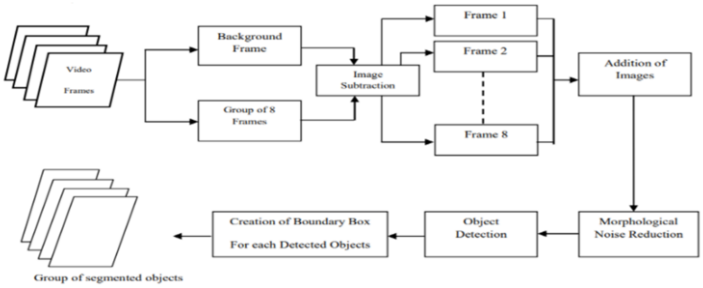

Figure 1. Image preprocessing and moving object detection process.

3. Sports Teaching Action Recognition Based on Hybrid CNN-HMM

3.1 Sports Information Acquisition

Combined with the actual use of chip functions, write the embedded software program of the system. The software design mainly includes four parts: software design of acquisition node, data processing, database design and control algorithm research. We have graphically configured a push-pull output mode pin in the STM32CubeMX software. In general, the data interaction between memory and IO devices is conducted through CPU, while DMA controller can realize the direct data exchange between memory and IO devices. It is necessary to determine a series of control parameters, such as the address, memory address and transmission direction of peripheral data. Before starting DMA transmission, DMA request should be issued. When a signal meeting the trigger condition enters the data acquisition card, the logic circuit (FPGA or PLD) on the board will drive the ADC to start sampling [6-8]. The timing function of the universal timer is enabled during programming. The flowchart of this scheme is shown in figure 2.

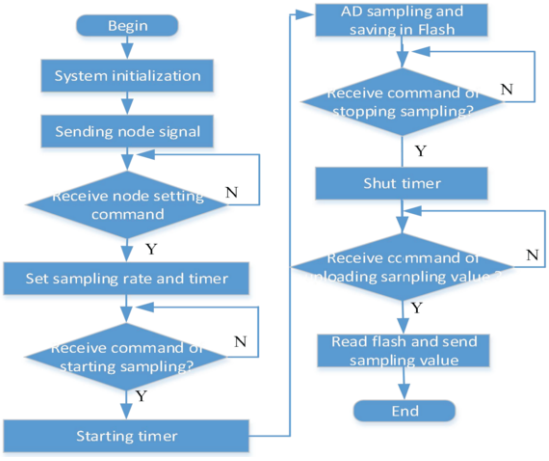

Figure 2 Specific process of acquisition node.

3.2 Image Extraction

The foreground image is calculated by calculating the difference between the current frame and the background model, but the focus is on how to build the background model. The parameter model of the background is used to approximate the pixel value of the background image, and the current frame is compared with the background image to detect the moving area. The pixel area with greater difference is considered as the moving area, while the pixel area with less difference is considered as the background area. In this paper, the median method is used to build the background model of video image. In a period of time, take a continuous N frame image sequence, arrange the gray values of the pixels at the corresponding positions in the N frame image sequence from small to large, and then take the middle value as the gray value of the corresponding pixels in the background image. The specific operation process is depicted as follows: First, set up a walking motion video sequence consisting of $1,2,...,n$ frames of images, then collect the gray value of the point at the same position as (x,y) in this image to obtain an array sequence: where i represents the image number. Therefore, the pixel value of the background image corresponding to this point can be represented by the middle value of the n frame image pixel value sequence,

$$B(x,y) = Median(P(x,y)) \tag{1}$$

where $B(x,y)$ is the pixel value of background image at point (x,y), (x,y) is the location of the pixel point. The extracted background image using the feature database is shown as figure 3.

Figure 3. Result of constructing background image by median method.

3.3 Action Recognition Using Hybrid CNN-HMM

In action recognition, the basic idea of hybrid CNN HMM is to use CNN to extract features from video frames and pass these feature sequences as input to the HMM model for action classification and recognition CNN can extract rich spatial features from video frames, while HMM can capture temporal information in action sequences. The advantage of hybrid CNN HMM lies in its ability to combine the excellent feature extraction ability of CNN with the sequence modeling ability of HMM [9].

(1) CNN is used to extract spatial features from video frames. Select CNN as the baseline classification algorithm and bring the segmented data segments into the hybrid CNN HMM algorithm for training. By conducting supervised learning on the training set, CNN can learn feature representations of different actions.

According to the characteristics of human behavior, we use the left right hidden Markov model to recognize human behavior. The data with the same action is obtained, and simple reordering of data with different groups but the same characteristics in the

time series is performed according to the characteristics. Learn the parameters of the same feature data at different times. If the data is large, we can conduct down sampling, which is similar to dividing the data equally in code. It can also be understood as the distribution of the same feature in different groups of data and in different time periods. Using HMM to recognize human behavior mainly involves the following three processes [10].

(2) The extracted feature sequence is input into the HMM model. To ensure the smoothness of the time series, the CNN classification results are combined with HMM, and in this step, the initial probability and transition probability of HMM are generated. In this case, HMM is used to model the transformation relationships between different actions and classify actions based on the observed feature sequences.

Determination of initial parameters. The initial HMM parameters are $\lambda = \{\pi, A, B\}$. First calculate the forward variable A and backward variables B. Then calculate the expectation according to the formula just introduced. In this paper, the initial state probability matrix is set according to the structural characteristics of the left and right HMM without crossing $\pi = (1,0,0,...,0)_{1\times N}$. The initial value of A is:

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 & \cdots & 0 \\ 0 & a_{22} & a_{23} & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN} \end{pmatrix} \tag{2}$$

where the sum of each line in the state probability transition matrix is 1. According to the feature of human activity we set $a_{n-1n-1}a_{n-1n}$, $n = 2,3,...,N$.

Since the size of human behavior codebook is set to M, that is, B is the state output matrix with the size of N rows and M columns. These optimal paths correspond to the possible values of the last implicit state value of each step. We set the initial value of observation probability matrix B as

$$b_j(k) = \frac{1}{M} \tag{3}$$

where $b_j(k)$ is the element at line j and row k in B.

(3) Training of parameters when the initial model is determined, the model parameters are trained using the known sequence of human behavior observations. According to EM algorithm, we need to write Q function first. Q is the logarithmic likelihood function of complete data. The expectation of conditional probability distribution of hidden variables under the premise of given model parameters and observation variables is depicted as follows:

$$Q(\lambda, \bar{\lambda}) = \sum \log P(O, I \mid \lambda) P(I, \mid O, \bar{\lambda}) \tag{4}$$

The observation symbol sequence $O = o_1 o_2 \ldots o_T$ is input to HMM model λ_i that is trained. By preceding algorithm, the output probability of O under λ_i is $P(O \mid \lambda_i)$, and the corresponding behavior of HMM to the maximum probability is the category of the sample to be identified.

Embed the posterior probability distribution results output by the CNN into the observation probability matrix of HMM, and use the Viterbi algorithm to reclassify them to obtain the optimal action sequence.

4. Simulations

The experiment selected the tennis players' decomposition action database as the research database, including 6 types of sports behaviors: walking, jogging, running, hitting, swinging and waiting. Each type of behavior is performed by 50 people in four different scenes (outdoor, outdoor with scale change, outdoor with clothing change, indoor) for many times. The visual angle of the video is fixed, and the background is relatively simple. Only one person does the action in each frame. In addition to category tags, the calibration data in the database also includes the silhouette of the actor in the foreground and the background sequence for background extraction. The video sample resolution is 1024×768, the frame rate is 20 fps (frame/second), and each subject collects 4845 frames. Two FACS experts marked the start and end of 12 kinds of AU on each frame of image in the database. In addition, it also contains a large number of interference pictures. The partial decomposition action image of the database is shown in figure 4.

Figure 4. Result of constructing background image by median method

In the identification phase, it is easy to judge whether the current running environment is running in virtualization mitigation according to the division of different attributes First, according to experience, CNN-HMM with N=4 and M=30 is used to identify five different types of behaviors: walking, running, hitting, waiting and squatting. Clip the video to the desired video clip and proportion through PR, and the sequence length is no less than 60 frames. It is not completely convinced that the time series filling which is purely dependent on the algorithm and cannot provide an explanation mechanism is insufficient, because the missing itself indicates that the sample information is insufficient. Figure 5 depicts the contour map of the centroid characteristics of the four people's jumping and bending movements after median

recognition.

Figure 5. Result of constructing background image by median method

Through the collection of samples, we can directly know which data results are positive and which results are negative. At the same time, we use sample data to run the results of the classification model. Table 1 shows the confusion matrix of the dictionary learning method for the database. Through the classification model, the prediction results and real attributes we get can be displayed in the form of a list, and it is found that the recognition results of tennis teaching decomposition actions in this method are the least confused, and the degree of confusion is higher than that of traditional recognition methods.

Table 1. Confusion matrix of teaching action recognition

	Walk	Run	Batting	Squat	Waiting
Walk	100	0	0	0	0
Run	0	100	0	0	0
Batting	0	0	93.4	0	4.2
Squat	0	0	3.6	100	2.8
Waiting	0	0	0	3.6	95.8

To test the distribution of input action sequences more intuitively and the differences in the results of action recognition experiments between the two methods, Table 2 shows the comparison between the actual input action sequences in the experiment and the action sequences recognized using single CNN and mixed CNN HMM methods. It can be seen that compared to a single CNN method, the difference between the recognition results of the hybrid CNN HMM method in the three experiments and the actual input action sequences is smaller, and the recognition accuracy is higher. From another perspective, the feasibility and superiority of the proposed method in human action recognition schemes are verified.

Table 2. Comparison of performance between single CNN and mixed CNN_HMM methods

	Accuracy		**Sensitivity**		**Specificity**	
	CNN	CNN-HMM	CNN	CNN-HMM	CNN	CNN-HMM
Action 1	98.12	99.99	93.25	99.98	98.80	99.99
Action 2	96.01	96.67	94.15	95.12	99.12	99.99
Action 3	95.1	96.65	98.80	99.34	94.18	97.35

Action 4	98.24	96.99	87.45	90.23	95.28	98.24
Action 5	97.49	98.53	74.13	87.12	98.98	99.99

5. Conclusion

According to the construction of the recognition system, this paper describes and applies it to physical education teaching and training, providing a scientific and intelligent training system for athletes and students. The identification system helps teachers observe in class and provides an intelligent research method. The program uses the process of information acquisition + image analysis + CNN-HMM classification to complete the collection, feature extraction, recognition and final classification of teaching actions. The experiment part takes tennis teaching action as an example, uses the samples in the real environment to obtain the action images of multiple student target students taught by the target teacher, and calculates the deviation value between the action and the standard, to formulate training plans to help them better complete the standardization of the action. The test results show that the auxiliary model has good recognition and classification ability, which is of great significance for training and selecting priority talents in sports field.

References

[1] Miao Xuelan. Sports action pattern recognition method based on fuzzy neural network theory. Computer Engineering and Application, 2000, 6:155-157.
[2] Warm, Wang Yifan. Artificial intelligence empowering sports: the application of computer vision in human motion recognition Journal of Shanghai Institute of Physical Education, 2020, 44(7): 25
[3] Panlili Elements and construction of motion recognition system based on sports visual image technology Bonding, 2020, 44(10): 91-93
[4] Haines D J, Kaiser K, Farrell A. How to Develop and Administer a College Recreational Sports Graduate Administrative Assistant Research Program. Recreational Sport Journal, 2009, 7:29-56
[5] Setiawan F, Yahya B N, Chun S J, et al. Sequential inter-hop graph convolution neural network (SIhGCN) for skeleton-based human action recognition. Expert Systems with Application, 2022, 6:195
[6] Zhu F, Zhu R. Dance Action Recognition and Pose Estimation Based on Deep Convolutional Neural Network. Traitement du Signal: signal image parole, 2021,2:3-8
[7] Wanni M. Action recognition model of athletes at the scene of the game based on SVM and multitarget tracking algorithm. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 2021(5):41
[8] Jiang J, Zhang Y. An Improved Action Recognition Network with Temporal Extraction and Feature Enhancement. IEEE Access, 2022, 10: 2-10
[9] Zhang Zhen, Zhang Shirong, Zhao Zhuanzhe, et al. Human Motion Recognition Method Using Hybrid CNN-HMM. Journal of University of Electronic Science and Technology of China, 2022, 51(3): 444-451.
[10] Avp A, Apa B, Iao A. Comparison of action recognition from video and IMUs. Procedia Computer Science, 2021, 186:242-249