International Journal of Multimedia and Ubiquitous Engineering
Vol. 9, No. 10 (2014), pp. 297-306
http://dx.doi.org/10.14257/ijmue.2014.9.10.29

Research on Action Recognition of Player in Broadcast Sports Video

Gang Liu[1], Deming Zhang[2] and Hui Li[1]

[1]*Sports Department, Harbin Engineering University, Harbin 150000, China*
[2]*Heilongjiang University of Chinese Medicine, Harbin 150000, China*
Liuganglg2014@126.com

Abstract

Based on support vector machine (SVM) and analysis of optical flow, the paper presents a new method for recognizing player motions in broadcast sports video. The video often has problems like bad-quality image, non-static video cameras and low-resolution image of player. To address them, from the perspective of movement analysis and according to the spatial distribution features of optical flow field of tracked members, the grid classification method, a kind of local analysis idea, is used to extract descriptive characteristics of movement recognition. With different idea from the traditional flow analysis, the method regards optical flow vectors in the traced areas as a kind of spatial distribution information in the mode of mobility, improving robustness of optical flow features. With SVM as model classifier and the application of time-sequential voting strategy, the type of player actions is identified. Compared with existing recognition methods based on apparent characteristics, the proposed recognition technique which fetches and depends on motional descriptive feature achieves better recognition effects.

Keywords*: Object tracking, Action recognition, Support vector machine, sports video*

1. Introduction

A player's movement trails in the competition is portrayal of his sports action information. Their actions in the contest can be important semantic clues to the understanding of competition schedule, detection of the event and extraction of wonderful clips. As far as the finer levels of visual content description are concerned, trajectory information can be considered as coarse-grained depiction, a macroscopic representation of the whole event. Gestures can be taken as fine-grained depiction, a detailed presentation of game process [1-2]. In order to more elaborate analysis and understanding of video content, it's necessary to recognize sportsmen's movements. For most competition events, we can classify them into two types based on the structural features of broadcast video: time-constrained [3-4]: such as football and basketball games; this type is mostly a group project, crew members to win the game by making higher scores than rivals through their cooperation in give time frame; the airing video is structurally loose, with and without highlights as well; score-constrained: the racket game like tennis and badminton match; this type is commonly individual work, one player to win out by making required points prior to his contestant with personal skills. When watching the racket game, audiences concern not only player footwork in the field and rather their gestures like stroke actions [5-6]. Thus we can learn that competitor action recognition means a lot to the analysis of visual contents regarding the racket game as above [7].

With the racket game like tennis and badminton as object of study, we'll discern player shots from broadcast visual counter-fight in many rounds [8]. Based on local movement analysis idea and grid classification method, a new movement descriptive operator based on optical flow field is developed, *i.e.* grid-classification-based optical flow histogram. With

ISSN: 1975-0080 IJMUE

SVM as pattern classifier to categorize player' actions, together with the utilization of time-sequential voting strategy, competitors striking movements are confirmed, including left swing, right swing and upper swing of the racket [9-10].

2. Player Tracking and Adjustment

In a broadcast on the racket games, shooting the cameras are generally non stationary. Therefore, there is a lot of camera movement in racket game broadcast video. The camera motion can make the motion feature extraction, and cannot reflect the motion information of player. This paper proposes the action recognition method to get the regional of game players by moving tracking of object, and player image is adjusting in the area. In order to eliminate the camera motion in the video broadcast [11].

Has been proposed for players of racket game video tracking method is adopting technique based on the template matching. It is very sensitive based on tracking method of template matching noise on broadcast video, easy to produce the tracking error. Due to the continuous accumulation of tracking error, resulting in tracking results quickly away from the real target area. The so-called "tracking drift" problem. Therefore, based on the tracking method of template matching cannot effectively track long time in the video. This can be made from [12] track strategy to prove. In [12], first broadcast tennis video is sequence segmentation in seconds, at 30 frames per second video sequence independently is player tracking based on template matching

In order to produce human adjustment image, the member area is tracking in horizontal direction and the vertical direction of regional expansion symmetry. Using the formula1 to Calculate expansion area "centroid" coordinate (m_x, m_y), the regional center is moved to the "centroid" coordinates, thus completing the generating of human adjustment image.

$$m_x = \frac{\sum_{x\in R}\sum_{y\in R} x\cdot f(x,y)}{\sum_{x\in R}\sum_{y\in R} f(x,y)}$$
$$m_y = \frac{\sum_{x\in R}\sum_{y\in R} y\cdot f(x,y)}{\sum_{x\in R}\sum_{y\in R} f(x,y)} \quad (1)$$

Which, R is tracking player area, $f(x,y)$ represents the gray values of the pixel in the $R(x,y)$.

After the above treatment, human adjustment image sequence corresponds to the camera always follow the player to shoot video results. Therefore, adjustment image sequence contains only caused movement by player limbs and racket. There is no camera motion video of the original broadcast. Figure1 shows some tracking and typical results of after adjustment. The figure1 is surrounded by a red rectangle area to get player adjustment image.

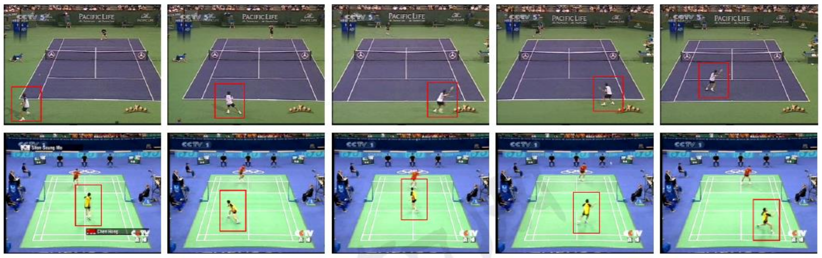


Figure 1. Results of Player Tracking and Stabilization

3. Computation of Motion Descriptor

Based on the movement analysis methodology, the proposed method creates descriptors to express different types of movements after calculating optical flow features of player adjusted time sequences. The key point to the use of optical flow features is such features that are obtained from abundant noisy broadcast video are extremely inaccurate. In works which used optical flow features for visual analysis, optical flow was regarded as time-sequential displacement information of every pixel within the video frame, which imposed higher requirements for the precision of optical flow estimation. In order to make effective use of optical flow features from the massive number of noisy broadcast sports video, we start the work with analysis of optical flow field, considering the vector field formed by optical flow vectors as a kind of movable spatial distribution information. Then, with the help of motional descriptors for the compact expression, the robustness of those features is enhanced. On the other hand, in the adjustment image sequence, camera movements are cleared. Movements in those images are as a result of relative displacement of player limbs and trunks. Those relative movements are shown in different areas in the adjustment image. Such local features can't be effectively expressed through global ones. Till now, we can take local analysis technique which bases on grid classification to divide the optical flow filed of adjustment image into non-overlapping subfields. Each subfield is called a grid. Through histogram statistics in each grid, the expression can be fulfilled of the spatial distribution model of optical flow field.

3.1. Optical Flow Estimation and Noise Erasure

We hope the acquired optical flow features of Adjustment image will only reflect player, the motion information of the image foreground. Hence, the background in the adjustment image will affect a lot the computation of optical flow. It's necessity to clear away background. Considering the background is sports court, we need to take the court modeling approach based on Gaussian mixture model and utilize region-growing algorithm for post-processing as to get full-view foreground player [13]. The procedure is displayed in Figure 2.

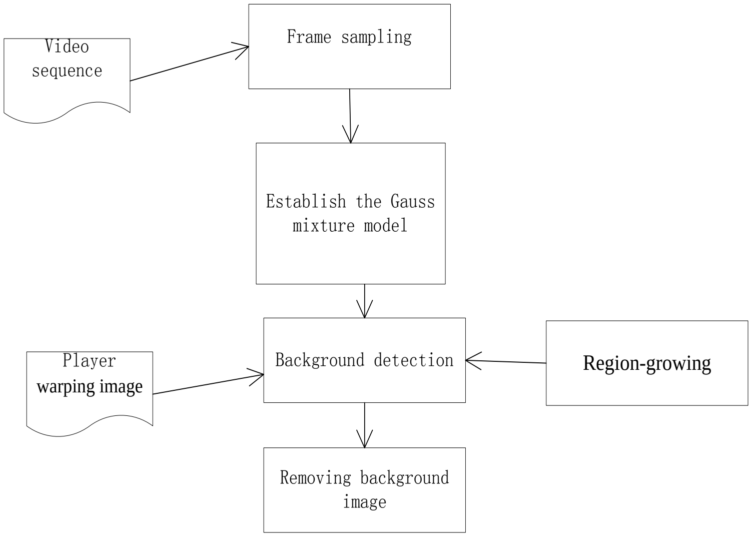


Figure 2. Background Subtraction for Human-centric Figure

The optical field can be estimated by according to player adjustment image sequences with clearance of backgrounds. However on account of the following points, the difference image of adjustment image of neighboring members is used to calculate the optical field. Firstly, player members' adjustment image will differ in terms of brightness for the change of camera's flashing and light intensity in the stadium. Such differences will lead to erroneous computation of the optical flow. Image differential calculation can remove impacts by the changing brightness. Secondly, academic findings on biological vision system suggest that human's visual cells are more sensitive to motional direction and velocity of object edges. Thus the estimation of optical flow based on image differences can better reflect the mechanism of human's vision system reacting against object movement. Based on the difference image, we take Horn-Schunck algorithm to estimate the optical flow field of player' adjustment image. The whole process of calculation can be formalized as:

$$\begin{aligned} DI_i &= HC_i - HC_{i-1} \\ OFF_i &= HS(DI_i), i = 2, \ldots, N \end{aligned} \qquad (2)$$

Which, DI_i is the adjustment image HC_i and HC_{i-1} difference images. $HS(\cdot)$ is SHorn-Schunck algorithm. OFF_i is the calculated optical flow field. N is the number of image sequence alignment.

3.2. Local Movement Analysis

As mentioned above, player' movements in adjustment images are caused by the relative displacement in their bodies, which are present in corresponding image zones. For different gestures, the spatial distribution of optical flow fields varies from one another, as seen in Figure 3. Regarding the upper swing of racket, optical flow vectors are densely distributed on the upper part in the adjustment image. In the normal image of left swing, the optical flow field on the left is more concentrated than the right. In contrast, in the adjustment image of right swing, the right is denser than the left. The proposed action recognition method just uses the distribution feature, making representation of it on the basis of local analysis idea. In the paper, it puts forward an effective region partition method, called grid division. As shown in Figure 4, optical flow fields after smoothing and de-noising treatment are divided into mutually disjoint grid region in the vertical and horizontal direction of adjustment image. On the theoretical sense, grid division can be structured in any arbitrary spatial forms, to be specific, the number of grids both vertically and horizontally is random. But considering player body structure in the adjustment image and computational complexity, we adopted 3×3 to divide grids. Too simple partition model can't give full description of the spatial distribution of optical flow; while too complicated partition like 5×5 and 7×7 will scale down grid areas and hence the histogram for estimating the optical flow will become sparse.

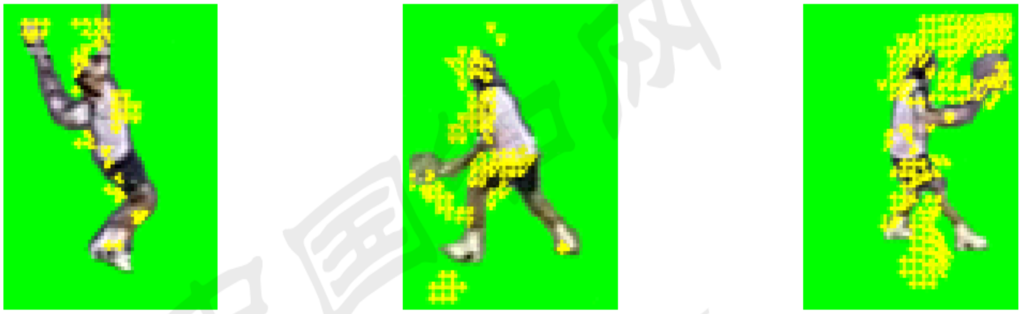
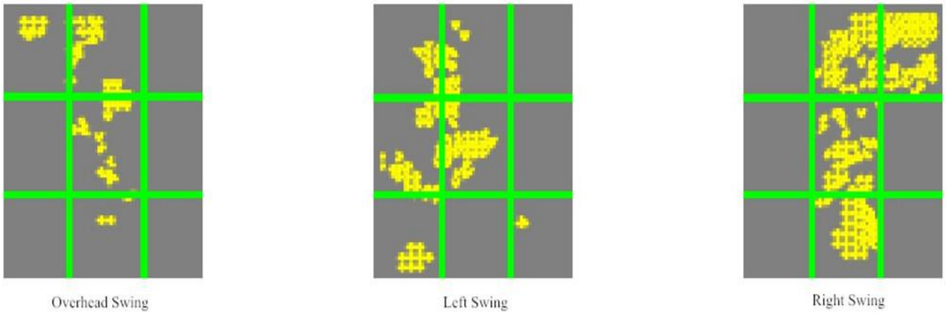


Figure 3. Spatial Distribution of Optical Flow Field for Player Actions

Based on kernel density estimation [14] for color layout and grid histogram for directional gradient [15], the paper developed Grid Based Optical Flow Histograms (G-OFHs) as the movement descriptor for player' swing of the racket. For the optical vector at coordinate p in the given optical flow field, its horizontal and vertical component is respectively $G_x(P)$ and $G_y(P)$. Then, we can define the amplitude $M(P)$ and directional angle $\theta(P)$ as follows:

$$M(p) = \sqrt{G_x^2(p) + G_y^2(p)}$$
$$\theta(p) = \arctan \frac{G_y(p)}{G_x(p)} \qquad (3)$$

Based on optical flow histogram grid division core idea: for each grid area, the arbitrary coordinate p optical flow vector is according to the amplitude of the $M(p)$ weighted quantization to the angle $\theta(p)$. Quantization weighting strategy not only considers their own magnitude $M(p)$, and using the method of kernel density estimation to consider distribution information of the adjacent optical flow vector.


Overhead Swing

Left Swing

Right Swing

Figure 4. Grid Partition for Optical Flow Field

4. Classification of Actions based on Supervised Learning

Here the recognition of actions is formalized into problem of classification, with the use of supervised learning algorithms to train classifiers for different gestures. Support vector machine is a very efficient supervised learning method. It has been widely applied for pattern recognition and classification. Compared with other supervised classification method, SV sorting algorithm proves powerful generalized ability to unknown testing data and it requires fewer setting parameters. Eight grid optical flow histograms which bases on adjustment image constitute input features of SVM classifiers, adopting Radial Basis Function $K(x, y) = \exp(-\lambda \cdot \|x - y\|)$ as kernel function to convert training samples from input space to hi-dimensional linear separable space. For the specified action sequences, with recognition results about actions and gestures in the frame and time-sequential voting strategy, player action included in the sequence are classified into left swing, right swing and upper swing of the racket.

In order to locate action sequence from the broadcast video, the two approaches are taken: the method based on audio keywords and the one based on analysis of player trails. With regards to racket competition events like tennis, for the reason of ball's heavy weight and player striking balls powerfully, the hitting sound is heard clearly and apparently in the video.

For such games, we can use audio keyword technique to test shot sounds as to navigate motional sequence in the video. We'll introduce that technique in the following section. As seen from Figure5, the video frame relating to the time point where the sound appears is named Hitting Point in the video sequence. Adjacent to those Hitting Point, we define a window W in the opposite direction to time axis. Thus, the video sequence formed by those video frames falling into window W is considered action sequence. Through observation of massive data, a swing of the racket by a player is basically completed in one second. Then, we set the length of window W the frame rate of sports video, such as 25 frames in PAL and 30 frames in NTSC.

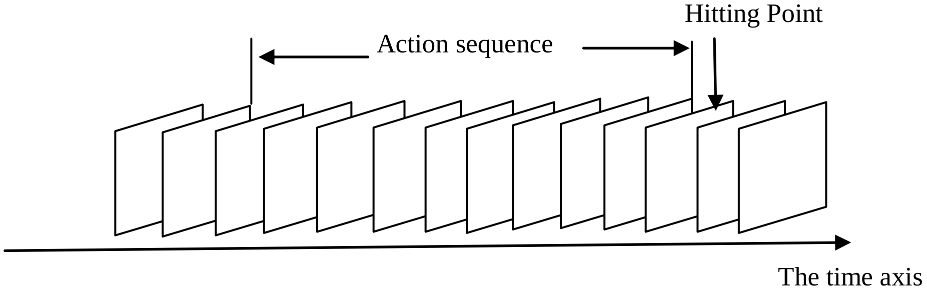


Figure 5. Location of Player Action Clip using Audio Keyword

5. Experiment Design and Discussion

Tennis and badminton are typical events among racket games. To validate the performance of the proposed method, we used three tennis matches like respectively Pacific Open Tennis 2011, French Open Tennis 2012 and Australian Open Tennis 2012 which were transcribed from broadcasting and TV program and the badminton match of the 2012 London Olympic game. Video were stored in MPEG-2 compressed format. The size of video frames is 352×288. The type of different swing actions in those matches was marked in manual mode. True values were created for Ground Truth, as listed in Table1, which includes match name, player, time length of the video and number of three swing actions.

Table 1. Experimental Data of Player Action Recognition

Match	Match Video	Number of Overhead Swing	Number of Left Swing	Number of Right Swing
tennis	Pacific Open 2011	41	91	136
	French Open 2012	249	689	278
	Australian Open2012	207	357	678
badminton	London Olympics2012	227	136	189
all	---	745	1189	1225

To appraise the performance quantitatively, we defined three qualifying indicators. Firstly, we defined recall rate (R) and precision rate (P) to confirm the identification ability of each gesture. The recall and precision rate is defined:

$$R = \frac{n_c}{n_c + n_m} \times 100\% \qquad P = \frac{n_c}{n_c + n_f} \times 100\% \qquad (4)$$

Which, for any action, n_c refers to the number of actions which are discerned correctly; n_m is the number of actions not be identified; n_f stands for the number of actions which are falsely recognized. Besides, we define accuracy (A) to evaluate the performance of recognizing (three actions) mentioned above. The definition is:

$$A = \frac{n_{c-overhead} + n_{c-left} + n_{c-right}}{n_{total}} \times 100\% \qquad (5)$$

Which, $n_{c-overhead}$ n_{c-left} $n_{c-right}$ indicates the number of correct recognition on the left and right swing. n_{total} is the number of all the action.

For tennis and badminton, we trained two classifiers to judge player movements. All data used three-times cross validation strategy to form training set and testing set, *i.e.* 2/3 data used as training sample and the rest as testing sample. After three iterative tests, the mean value of three evaluation data was considered as the final result. Table 2 shows the experimental results of the algorithm.

Table 2. Experimental Results of Player Action Recognition

Match	Action categories	Number of action	Recall rate (%)	Precision rate (%)	Accuracy (%)
tennis	Overhead Swing	499	90.4	93.2	90.7
	Left Swing	1047	87.8	89.9	
	Right Swing	1051	93.8	91.7	
badminton	Overhead Swing	255	87.6	85.3	87.6
	Left Swing	136	85.3	91.3	
	Right Swing	170	89.4	92.7	
all	---	3158	---	---	90.2

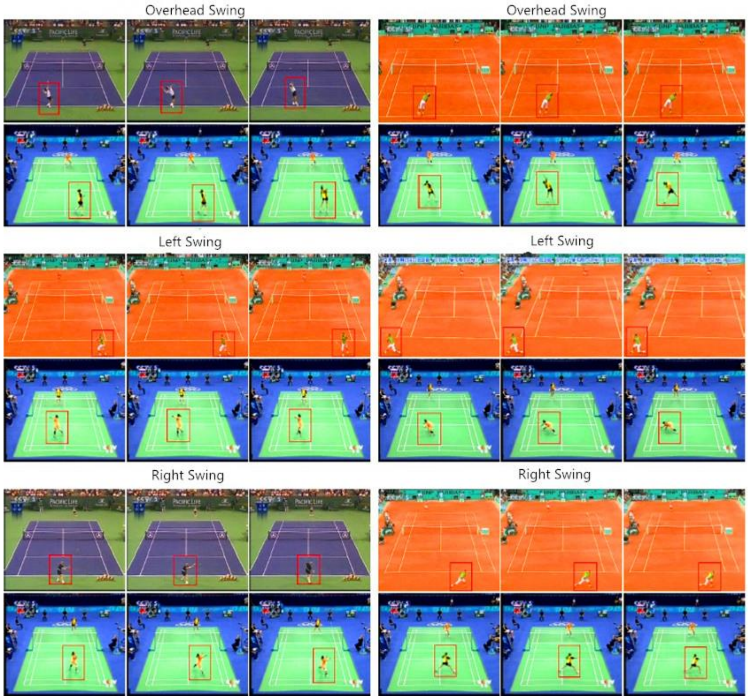
Obviously from table2, for tennis and badminton games, the paper proposed method had good accuracy rate of recognition at separately 90.7% and 87.6%. For those three actions of racket swing, the accuracy rate is 90.2%. In the test video, player' images have 30-40 pixels. Due to factors like low-quality video, camera movements, actions of swinging are un-sharp in every detail. The above research findings confirm that the proposed motional descriptors and recognition strategy are greatly effective. Figure 6 (a) presents some featuring action sequences which were correctly identified. The typical sequences which were wrongly discerned are seen in Figure 6 (b). In the experiment, the wrong recognition happened because contestants had to take unusual gestures to hit the ball for highly difficult return or excessive consumption of physical strength. In other words, in the eyes of professionals like coach, player' actions changed. The optical flow field distribution of those abnormal shots is different from the normal. In consequence, the optical flow histogram based on grid division can't accurately portray the spatial movements of those actions.

While evaluating the action recognition method based on motional analysis, we compared the effectiveness between the proposed method and existing ones which base on the analysis of apparent characteristics. In the work [16], the author followed tracks of participants to obtain the area. After reduction of background to get the outlines of their posture, Karnunen-Loeve (KL) transform was utilized to map the information of outline into an Eigen-space. Then, eigenvalues are arranged in a descending order, with the Eigenvector of the first m biggest values to form apparent descriptors of actions and gestures. Finally the nearest neighbor classifier was used to recognize action sequences by according to those descriptors. Like the foresaid experiment, three-times cross validation strategy is applied to separate all data into training set and testing set. After three repetitive tests, the mean value of three evaluation results is deemed the final outcome. In the experiment, through modifying and choosing the percentage of apparent vectors after rank ordering, we simplified the structure of action descriptors. With the dataset used here, when the front 80% apparent vectors were chosen, the recognition achieved the best rate of accuracy. Table3 lists out appraisal results by the recognition algorithm through apparent analyses.

From Table 3, we can conclude that the proposed method is superior over the apparent analysis. In the broadcast video, for different angles of camera shooting or direction variation of player' movement, the profile information of their actions doesn't have good inter-class discrimination. Hence, the apparent descriptor based on profile restructuring doesn't have excellent classifying robustness. Through comparisons, the motional descriptor in the paper is much better than others in terms of robustness.

Table 3. Experimental Results of Player Action Recognition using Appearance Analysis

Match	Action categories	Number of action	Recall rate (%)	Precision rate (%)	Accuracy (%)
tennis	Overhead Swing	499	76.6	69	71.5
	Left Swing	1047	73.4	70.5	
	Right Swing	1051	67.1	75.4	
badminton	Overhead Swing	255	57.3	62.3	59.1
	Left Swing	136	64.0	75.0	
	Right Swing	170	57.6	66.7	
all	---	3158	---	---	69.4



(a) Action clips of correct recognition (b) Action clips of false recognition

Figure 6. Results of Representative Recognition for Three Swing Actions

6. Conclusion

This paper studies player action recognition problem in broadcast sports video. In view of the recognition method based on the analysis of apparent, proposed action recognition method based on movement analysis, the players swing is distinguishing in racket sports broadcasting video. In this paper, from the angle of motion analysis, based on optical flow characteristics presents a new low-resolution video human action recognition algorithm. The kernel density estimation method and direction of the gradient histogram grid combined to calculate the motion descriptor. Through the contrast experiment, action recognition algorithm can effectively identify of player action in broadcast sports video, improve the overall performance of the original algorithm.

References

[1] H. Miyamori and S. Iisaku, "Video Annotation for Content-based Retrieval Using Hu-man Behavior Analysis and Domain Knowledge", IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble: IEEE, (**2000**), pp. 320–325.

[2] H. Miyamori, "Improving Accuracy in Behavior Identification for Content-based Retrieval by Using Audio and Video Information", IEEE International Conference on Pattern Recognition, Quebec: IEEE, vol. 2, (**2002**), pp. 826–830.

[3] M. Roh, B. Christmas and J. Kittler, "Robust Player Gesture Spotting and Recognition in Low-resolution Sports Video", European Conference on Computer Vision, Graz: Springer, (**2006**), pp. 347–358.

[4] R. Li, L. Wang and K. Wang, "Research on human action recognition", Pattern recognition and artificial intelligence, vol. 1, (**2014**), pp. 35-48.

[5] P. Yingxu, "Chinese National Men's volleyball players training practice analysis", Journal of Capital Institute of Physical Education, vol. 1, (**2014**), pp. 65-69.

[6] R. Li, L. Wang and K. Wang, "Research on human action recognition", Pattern recognition and artificial intelligence, vol. 1, (**2014**), pp. 35-48.

[7] Y. Pan, "Chinese National Men's volleyball players training practice analysis", Journal of Capital Institute of Physical Education, vol. 1, (**2014**), pp. 65-69.

[8] Q. Wang and L. Xia, "The player behavior of basketball video China prediction", Journal of image and graphics, vol. 4, (**2012**), pp. 560-567.

[9] Y. Zhou and L. Wang, "The visual human action recognition", Journal of Shandong Light Industries College (NATURAL SCIENCE EDITION), vol. 1, (**2012)**, pp. 85-90.

[10] P. Qu, S. Qu, T. Kang and Y. Zhao, "Sports video intelligence in low-level visual information analysis", Journal of Sports Adult Education, vol. 3, (**2012**), pp. 49-51.

[11] Q. Lv and S. Li, "Research of human - interactive visual recognition method on. computer engineering and design", vol. 8, (**2012**), pp. 3194-3199.

[12] G. Sudhir, J. Lee and A. Jain, "Automatic Classification of Tennis Video for High-level Content-based Retrieval", IEEE International Workshop on Content-Based Access of Image and Video Databases. Bombay: IEEE, (**1998**), pp. 81–90.

[13] Q. Ye, W. Gao and W. Zeng, "Color Image Segmentation Using Density-based Clustering", IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong: IEEE, vol. 2, (**2003**), pp. 401–404.

[14] D. Comaniciu, V. Ramesh and P. Meer, "Kernel-based Object Tracking", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 25, no. 5, (**2003**), pp. 564–577.

[15] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", IEEE International Conference on Computer Vision and Pattern Recognition. Santa Diego: IEEE, vol. 1, (**2005**), pp. 886–893.

[16] H. Miyamori and S. Iisaku, "Video Annotation for Content-based Retrieval Using Human Behavior Analysis and Domain Knowledge", IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble: IEEE, (**2010**), pp. 320–325.

Author

Liu Gang, He is an Associate Professor at Sports Department of Harbin Engineering University. He is in the research of physical education and training