International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.9 (2016), pp.83-92
http://dx.doi.org/10.14257/ijmue.2016.11.9.09

Sports Video Structure Analysis and Feature Extraction in Long Jump Video

Maohua Zhuang

Harbin Institute of Sports, Harbin 150000, China
1023054464@qq.com

Abstract

How to help people to find their favorite sports in the massive video, in order to achieve this goal in the finite state machine (FSM) theory based on in-depth research. Combined with the actual situation of this paper, first, we determined the reasonable FSM model. Then, a fast robust global motion estimation algorithm is used to estimate the global motion of the video sequences, and the foreground is separated from the background by motion compensation. After the foreground of the video, a series of image feature extraction algorithm, which is based on the main color histogram and histogram, is used to extract the low level feature extraction. Finally, the results of the experimental operation, the performance of the algorithm, and the results of the system are presented.

Keywords*: Sports Video, Finite-state Machine, Field Knowledge, Long Jump Video*

1. Introduction

For video programs about sport competitions, TV program producers concluded a complete set of scientific and reasonable production and editing mode based on many years of experience in broadcasting sport matches [1-3]. Almost all videos of sport contests are produced by following that pattern, because only in that mode, can TV audiences enjoy watching sport games to the greatest extent [4-6]. The typical video program compiling model leads to typical sport competition video structure. That makes it possible for us to conduct semantic-based analysis, retrieval and query of sport videos. And thus it's a significant task to analyze the structural features of sport videos [7].

Like other video data, jumping video data are enormous. For the purpose of effective organization of related videos, it's required to decompose videos to elementary units. It's generally accepted that the basic physical unit of video is shot [8-9]. One shot is composed of numerous frame images which are acquired consecutively in time by one camera. The detection of shots is an issue of segmenting videos from the perspective of time domain. To edit shots in different ways, they can join up to form video programs. Different video programs have own uniqueness [10-11]. So it's necessary to use different methods of shot segmentation for various video programs, to realize the sound decomposition of videos. Then on that basis, further analysis is made in order to perform nonlinear browsing and semantic-based inquiry and retrieval [12-13].

Finite state machine (FSM) is a mathematical model to describe the complex system by simplifying assumption. In various UML tools, finite state machine (FSM) is a powerful tool for supporting dynamic modeling. Through a chart, it can describe a complex logic, which can effectively support the modeling of complex behavior. It is widely used in communication protocols, graphical interfaces, and many other applications.

ISSN: 1975-0080 IJMUE

2. Analysis of the Structure of Jumping Videos

Generally, one segment of sport video data can be divided into several scenes, of which each contains one to more shots; one shot is a series of successive recording image frames, representing continuous action in one time frame or the same place. Shot is decided by the onset and ending of one-time shooting by a camera. The structure of one video scene is a succession of shots which are semantically associated, which occur at the same location and time and appears the same people or event. Hence video information can be classified to four hierarchical structure: video, scene, shot and image frame, from crudeness to fineness. Of them, frame is the smallest video data, a static picture; shot is basic unit of video data, a continuous action of camera; one shot covers what happening continuously in the nearby location; scene is formed by shots with similar contents, describing one event from different angles; the whole video is consisted of plentiful scenes, telling a complete story. It is shown in Figure1.

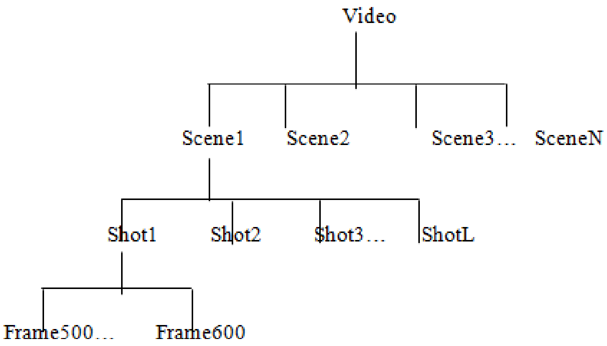


Figure 1. Video Structure Diagram

Any video is linked up by shots one after another, shot transition from one to another. Due to different ways of linking, shot change has abrupt change and gradual change. If one shot is transferred suddenly and directly to another one. It's abrupt change; gradual change is a gradual transition from one shot to another, no obvious shot hop, including fade-in, fade-out, transition etc. On the part of video compilation, gradual change is achieved by editing colors and space. Color editing is the treatment of video frame colors which transit gradually to the next shot, such as dissolving, fade-down, perspective. Spatial editing is the technique of adjusting the spatial position of video frames to gradually switch to the next shot. During camera shooting, according to different requirements, shots in different moving states can be obtained by using different camera moving ways to cope with shots.

3. Studies on Methods for Detecting Jumping Sport Video Shot Boundaries

The accurate detection of shot boundaries is foundation to make subsequent detection with FSM, which can be considered as pre-treatment. Shot boundary detection is discussed for too many years. Many reliable approaches were proposed to detect abrupt change shots, which, however, proved defective, for example, when fast-moving object in the video frame, or for the explosive scene, the existing detection methods are not effective. The detection methods for gradual change shot need improvements as well. For sport videos, shot boundary detection problem has not been solved, for the reasons as follows:

(1) Unlike common videos, in sport videos, the camera focuses always on the competition field, which is often monochromatic, such as football pitch; colors of

competition field is primary color among frames; strong color association exists between frames; frame color histograms of different shots do not change much and that traditional detection methods become futile with the utilization of inter-frame histogram difference.

(2) In sport videos, object's motion intensity in frame images is very big. To trace high-speed moving objects, shooting techniques like panning and zooming are usually employed. Some conventional ways can hardly find out shot boundaries.

(3) In sport videos, a certain amount of gradual shots is contained. Just as mentioned previously, no accurate detection approach is presented to deal with such complicated shots.

So a reliable shot boundary detection method is a critical topic. Ahmet Ekin *et al*. suggested a football video shot segmentation solution based on multiple thresholds; it's difficult to decide the appropriate threshold value for the method; besides, semantic content of each segmented shot is not quite definite. After reviewing domestic and foreign achievements by plenty of researchers, we use one method to extract shot content changes through foreground/background segmentation based on the global movement and color histogram variation of adjacent frame sequences. It is shown in Figure2.

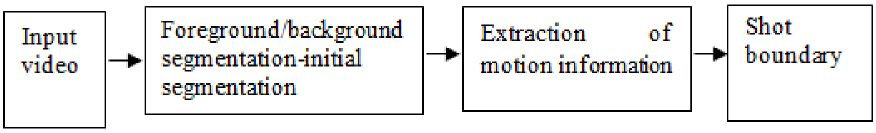

Input video
Foreground/background segmentation-initial segmentation
Extraction of motion information
Shot boundary

Figure 2. Shot Segmentation Method Diagram

3.1. Foreground/Background Segmentation-Initial Segmentation

In sport action video images, there're two kinds of motion: overall motion, *i.e.* background movement caused by camera motion; partial motion, *i.e.* foreground movement caused by sportsmen. The acquisition of accurate global motion parameter is key and foundation to foreground and background separation and sportsman body extraction and movement analysis. Here we use global motion estimation to execute foreground and background split of video frames. Considering characteristics of sport motion videos, the global motion of background caused by camera movement is expressed by parameter affine motion model as:

$$\begin{cases} x = ax'+by'+e \\ y = cx'+dy'+f \end{cases} \tag{1}$$

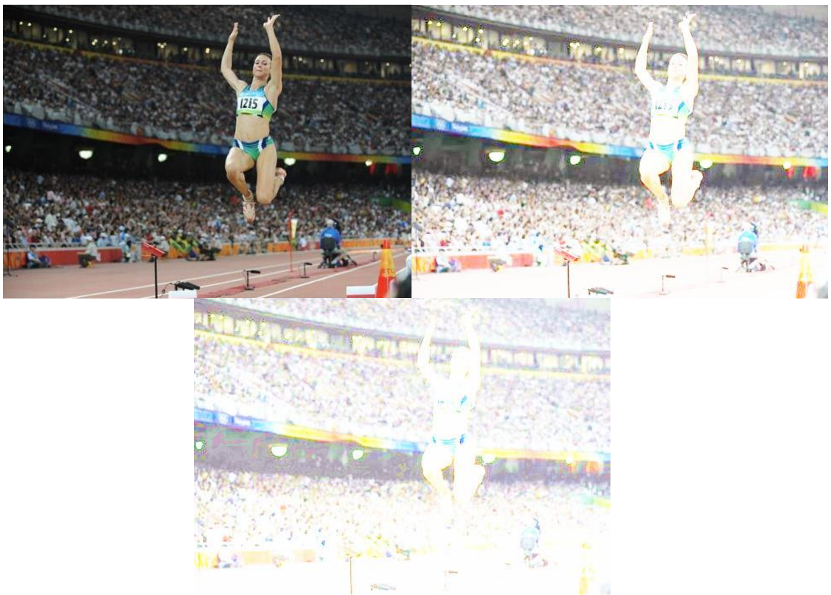
Current video motion object extraction methods are simply of two types: one based on sequential attribute, cutting motion objects as per video sequential property [14-15]; the other based on spatial attribute, segmenting motion objects as per image zone or edge information. However, no matter which is used to segment motion objects, visible background and irregular movements of objects will lead to reduction of segmentation accuracy, because both methods split up motion objects and background area by means of motion information [16-18]. But in movement analysis, static foreground area is easily falsely detected as foreground or background due to manifested background and object's irregular movements, downgrading the precision of segmentation.

We introduce a new method for the fetch of moving objects based on dynamic background construction. Firstly, the dynamic background construction technology based on foreground separation makes use of multi-frame differences to construct the present background; then by background subtraction, it splits out motional object as to remove noticeable background in segmentation result. Meanwhile, it detects static foreground area as per sequential information and merges it to target area as to get the complete object area, overcoming the impacts of object's random movements on segmentation accuracy. Finally, regarding edges of object area as initial position, it applies active contour model which uses color gradient as external energy to get precise profile of moving objects.

Let I_k for the current frame, $I_i(i=k-L,...,k+L)$ is $2L+1$ frame continuous image, the global motion parameters between adjacent frames are $\theta_{k-L+1},...\theta_{k+L,k+L-1}$. In order to construct I_k background. To calculate the parameters θ_i of spatial coordinate theta is aligned to the I_k on I_i.

$$\begin{cases}\boldsymbol{\theta}_k = (1,0,0,1,0,0) \\ \boldsymbol{\theta}_i = \boldsymbol{\theta}_{i+1,i}\cdot\boldsymbol{\theta}_{i+1} & ifi<k \\ \boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1}\cdot\boldsymbol{\theta}^{-1}_{i,i-1} & ifi>k\end{cases} \tag{2}$$

In order to obtain the accurate and robust global motion parameters, we propose two step method for estimating the global motion parameters: Used the MPEG-4 algorithm, the Konrad algorithm [19] is used to get the preliminary estimation results; After current frame background is constructed, it's possible to rapidly divide out motion object area by eliminating construction background from current frames and binarizing it. We utilize significance testing technology to binarize frame differences after background subtraction, since the method is less complicated and can help determine threshold value according to certain fault-tolerant rate. If one part of a moving object is still at one time range, we think partial characteristics of static foreground area (*i.e*. the area) may be left in constructed background, no obvious difference between the area and background. As a result the area can't be obtained by background removal. In this case, we use time-series relationship to detect inter-frame static foreground area and combine it with object area got by background elimination as to have the complete moving object region. It is shown in Figure3.



(a) Original Image (b) Using Konrad Algorithm to Deal with the Background Map (c) Results of Differential Image bBnarization

Figure 3. Complete Moving Object Region

3.2. Extraction of Motion Information Further Segmentation

Color histogram is simple and effective so that it's widely applied for content-based image retrieval. Naturally, color histogram gains wide use in content-based video retrieval. Paper [20] introduced a color histogram which is called Alpha-trimmed mean histogram, including mean histogram and median histogram. If a color histogram is fetched for every frame image, then a group of image frames will have a set of color histograms. To merge them to one color histogram, we need to sort out the value of same histogram grids, deleting the highest and lowest values and take an average. It takes similar ideas as the scoring scheme used for singing contest. We take for instance, 10 referees score a singer, removing two highest and two lowest scores, then calculating the average value.

We propose to do by main color extraction and tracking and use main color histogram to describe a group of image frames. To describe such a histogram not only cuts the size of histogram and enhances the result of histogram matching, because the main color content is grasped and it's less affected by noises. The main color histogram for depicting a set of image frames considers the main color of a single image and time change of the main color. Such kind of representation method catches the nature of video as consecutive time-based media. Unlike other histograms, main color histogram utilizes time information and takes some semantic considerations. A group of image frames is a general concept, which can be mirror, sub-shot or a group of shots. In the following passage, we regard one shot as a group of image frames and assume the shot includes unitary topic (*i.e.* content is consistent). Alternatively we divide the shot to a few sub-shots of consistent contents and use as one image frame group.

First of all, calculate color histogram of each frame image; then find out the main color of the one frame. The color model we chose to use here is HSV because according to Euclidean distance, colors are uniformly distributed in HSV color space; with three-dimensional Cartesion coordinate system for quantification; axis X and Y quantified to 20 values and axis Z (brightness) to 10 values, in Figure 4. The pixel or DC block of frame I (when MPEG1/2 data used) is projected to quantized HSV color space. In the 3D color space, the distribution after normalization forms a normalized 3D color histogram. In it, we can discover all important local maximum points. We define the small ball which contains the local maximum point and whose diameter is 3 quantified units as one color object. In those color objects, the one with most pixels (we'll take first 20 colors in the experiment) as main color object in one frame. Note that they are not corresponding to a space object in image frames.

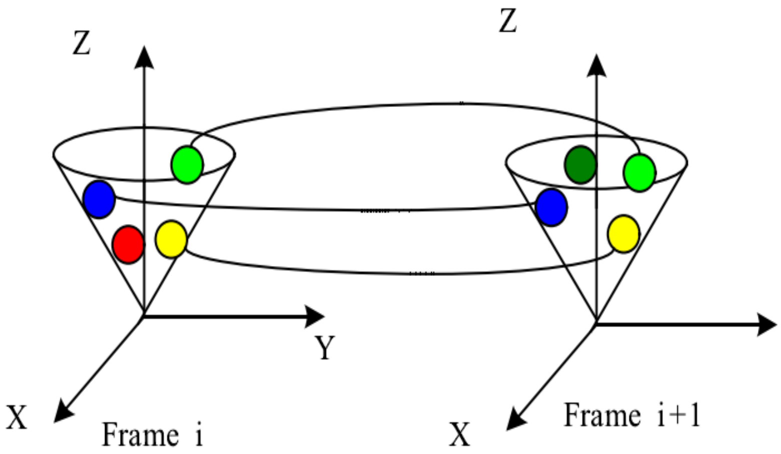


Figure 4. Main Color Extraction and Tracking

3.3. Feature Descriptor for Video Frames

The color block diagram obtained from the above calculation can obtain several distribution characteristics, including the area histogram Harea, position histogram Hpos, the regional variance histogram Hvx, Hvy, X and Y direction. In the X and Y directions, the region length and width histogram Hsx, Hsy. They are defined as

$$(1)\quad H_{area}(i)=\sum_{R_J\in\Omega_J}Area(R_j),i=0,1,2,...,7$$

$$\Omega_i=\{R_j\big|Area(R_j)\in(A_i,A_{i+1})\},i=0,1,2,...,7\qquad A_i=1/2^{8-i},i=1,2,...,8;A_0=0$$

$$(2)\ H_{pos}(i)=\sum_{R_j\in\Omega_i}Area(R_j),i=0,1,2,...,15$$

$$\Omega_i=\{R_j\big|Center(R_j)\in Block(i)\},i=0,1,2,...,15$$

$$(3)\ H_{vx}(i)=\sum_{R_J\in\Omega_I}Area(R_j),i=0,1,2,...,7$$

$$\Omega_i=\{R_j\big|\sigma_x(R_j)\in(B_i,B_{i+1})\},i=0,1,2,...,7\qquad B_i=1/2^{8-i},i=1,2,...,8;B_0=0$$

$$(4)\quad H_{sx}(i)=\sum_{R_J\in\Omega_I}Area(R_j),i=0,1,2,...,7$$

$$\Omega_i=\{R_j\big|Width(R_j)\in(B_i,B_{i+1})\},i=0,1,2,...,7\ B_i=1/2^{8-i},i=1,2,...,8;B_0=0$$

4. Experimental Analysis and Results

4.1. Environment of Simulation Experiment

4.1.1. Video Database

Experimental data was collected from television recording sport programs. The video database is very challenging. It lasts 2hours and forty five minutes, totaling 3514 shots and 226936 frame images, including advertisements, sport news, and some different sport competition program fragments. Some are similar video clips like news titles, advertisements; some are repetitive video clips like different car racing matches, same ads in different duration and editing.

4.1.2. Feature Library

The feature library includes two visual features like color and texture and high-level semantic features of motion information included in the moving objects. Low-level features like color are expressed by main color histogram and accumulative histogram, motion information of moving objects in key frames extracted as high-level semantic features of video sequence, which is further divided into video fragments with semantic concepts. To extract motion information, the global movement estimation Konrad algorithm and exterior point filtering algorithm based on Fisher linear discriminant criteria.

4.1.3. Inquiry Mode

The inquiry based on FSM is FSM template established respectively for various sport competition. When query request is sent, inputting video clips to according sport competition FSM will help user find interesting sport program.

4.1.4. Matching Method

Used maximum matching and optimal matching to achieve similarity measurement of video clips

4.1.5. Evaluation Indictors

Choose images with target as a group of relative images; then calculate recall and precision ratio based on return results; the higher the recall and precision ratio is, the better performance the retrieval algorithm realizes.

4.2. System Implementation

Query interface is an important way for man-machine interaction. A good query interface can facilitate users to acquire various information without too many obstacles. In designing the interface, it requires to consider fully requirements of different users and their habits and preferences. Therefore, how to provide a simple and friendly interface and implement rapid retrieval of images is another important concern and topic in current days. The system is windows 2000 and completes in visual c++ 6.0 development environment. We cut from diversified sport matches out video fragments to constitute two video datasets, one for training parameters and the other for validating the model effect. The whole system is consisted of two modules: retrieval module and discriminant module. Through the system, users provide initial video clip query request to the system, retrieving easily and quickly the desired sport competition video clips.

4.3. Experimental Results and Performance Evaluation

4.3.1. Retrieval of Accurate Clips

From Table 1, we see that the proposed method and that in [21] achieved high recall rate; but the method here obtained better precision ratio than [21], because [21] considers only the number of two clips of similar shots while the algorithm takes into account the corresponding relation of similar shots. In terms of retrieval speed, our method is quicker than [21]. According to the experiment, total retrieval time is equal to similar shot discriminant time.

4.3.2. Retrieval of Similar Clips

In Table 2, either recall or precision ratio, the proposed algorithm is better than [21]. Inquiring fragment 1 and 2 are too complicated. In the video database, jumping competition appeared four times. We lost two of them because we used blue runway for the query, one of which is green color and the other relates to shots of competitors and audience, with few shots of blue runway. Similar to inquiry clip 1, clip 2 is strongly semantic and can hardly be utilized. The whole clip reflects basic color characteristics of the semantics. The method here made satisfactory retrieval effects. It indicates that the query of video clip formed by a few shots can have better effect than single shot or image query. Also the method did quicker than [21]. The longer the query fragment is, the better our method proves. For example, when searching clip 2, the proposed method did quicker than [21] by over five times. Another advantage is the approach in the paper arranged similar fragments from big to small similarity. Apart from the visual feature, similar fragments with different factors were considered. Contrarily, the similarity in [21] depended only on the quantity of similar shots. Through testing on several persons, the proposed solution proved to be more accordant with human visual and psychological features in the ranking of similar video clips.

Table 1. Retrieval Results of Accurate Clips

Query clip	Frame numbers	The proposed approach			Match and tiling approach		
		Precision (%)	Recall (%)	Speed (s)	Precision (%)	Recall (%)	Speed (s)
Long jump	836	88	92	108	75	90	240
High jump	725	87	88	74	84	84	198
Shot-put	375	85	87	89	81	78	103
Run	554	79	86	109	82	86	149

Table 2. Retrieval Results of Similar Clips

Query clip	Frame numbers	The proposed approach			Match and tiling approach		
		Precision (%)	Recall (%)	Speed (s)	Precision (%)	Recall (%)	Speed (s)
Long jump	507	78.3	50.6	48	77.1	50.3	145
High jump	378	82.5	87	118	50.9	50.7	509
Shot-put	378	87.3	86	89	82.6	50.9	99
Run	544	84.5	89	107	81.6	88.9	148

5. Conclusion

In this paper, we first introduce the data structure of sports video, and the video data can be divided into four levels: video, Scene, Shot and Frame. Then, the research results of other researchers are presented, and the segmentation method of motion scene recognition based on global motion is proposed. The features of the frame image extracted from the shot segmentation process can be used for subsequent calculations. Finally, the distribution of several color block diagrams is obtained by the color block diagram of the video frame.

References

[1] Y. Lelin, “Research and analysis of video semantic retrieval based on visual information”, Beijing University of Posts and Telecommunications, **(2012)**.

[2] “Research on video structure analysis of cosmos and automatic cataloguing technology”, Beijing University of Posts and Telecommunications, **(2013)**.

[3] L. Xiaowei, “Sports video analysis and customization”, Wuhan University of Technology, **(2010)**.

[4] L. Qingshan, T. Xiaofeng and L. Hanqing, “Analysis of sports video”, Chinese Journal of computer, vol. 7, **(2008)**, pp. 1242-1251.

[5] T. Jie and W. L. Ying, “Video summarization method based on feature animation”, The research and application of computer, vol. 10, **(2009)**, pp. 3960-3962.

[6] Q. Ping, S. G. Qu, K. Tao and Y. L. Zhao, “Based on underlying visual information of sports video intelligent analysis”, Journal of sports adult education, vol. 3, **(2012)**, pp. 49-51.

[7] Z. Y. Xiaobo and H. Cao, “Sports video annotation and indexing based on hierarchical semantics”, Computer application and software, vol. 10, **(2012)**, pp. 258-260.

[8] C. Ming, D. Liwei and J. J. Chan, “AAM and HCRF of the football video exciting event detection”, Computer research and development, vol. 1, **(2014)**, pp. 225-236.

[9] C. Yunxiang, “Sports video annotation based on the characteristics of 2D human joints”, Computer process, vol. 4, **(2014)**, pp. 252-257.

[10] C. Ming, D. Liwei and L. Yingying, "Multi-dimensional semantic clues and HCRF model of soccer video exciting event detection", Computer aided design and computer graphics, vol. 11, **(2013)**, pp. 1715-1724.

[11] X. Guoqiang and D. Yi, "Track and field based on sequential pattern mining", Micro computer information, vol. 3, **(2011)**, pp. 221-223.

[12] Z. Yingying, Zhuyanyan, the Zhenkun, "Sign lens and the bag of words model sports video classification based on types", Computer aided design and graphics journal, vol. 9, **(2013)**, pp. 1375-1383

[13] L. Xueying, L. Yun and H. Chao, "Analysis of sports video based on Dynamic Bayesian networks", Micro computer information, vol. 21, **(2010)**, pp. 9-10.

[14] T. Aach and A. Kaup, "Statistical model-based change detection in moving Video", Signal Processing, vol. 31, **(1993)**, pp. 165-180.

[15] J. Guo, J. Kim and C.-C. J. Kuo, "Fast and accurate moving object extraction technique for MPEG-4 object-based video coding", SPIE Visual Communication and Image Proceedings, San Jose, CA, **(1999)**.

[16] Y. Tsarg and A. Averbuch, "Automatic segmentation of moving objects in video sequences: a region labeling approach", IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 7, **(2002)**, pp. 597-612.

[17] I. Patras, E. A. Hendriks and R. L. Lagendijk, "Video segmentation by MAP labeling of watershed segments", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, **(2001)**, pp. 326-332.

[18] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking", IEEE Transactions on Circuits and Systems for Video Technology, vol. 8, no. 5, **(1998)**, pp. 539-546.

[19] F. Dufaux and J. Konrad, "Efficient, robust, and fast global motion estimation for video coding", IEEE Transactions on Image Processing, vol. 9, no. 3, **(2000)**, p. 85-88.

[20] A. M. Ferman, S. Krishnamachari, A. M. Tekalp, M. A. Mottaleb and R. Mehrotra, "Group-of-frames/pictures color histogram descriptors for multimedia applications", ICIP2000, **(2000)**.

[21] L. P. Chen and T. S. Chua, "A match and tiling approach to content-based video retrieval", In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2001), **(2011)**, pp. 417-420.

Author

Maohua Zhuang, She received her B.S degree from Harbin Institute of Sports. She is an Associate Professor at Harbin Institute of Sports. She is in the research of physical education and exercise training.