A New Method for Shot Identification in Basketball Video

Yun Liu
College of Information Science and Technology of
Qingdao University of Science and Technology, Shandong province, China
Email: Lyun-1027@163. com

Xueying Liu and Chao Huang
College of Information Science and Technology of
Qingdao University of Science and Technology, Shandong province, China
Email: { lxying2009, hchaopro }@qq. com

***Abstract*—This paper presents semantic-based shot event identification. Based on Dynamic Bayesian Network (DBN), the gap between low-lever features and high-lever semantic will be resolve. We apply the mean-shift algorithm and Kalman filter to identify and track the ball and SURF(Speed up Robust Features) to find the basketball hoop. At last the DBN is applied to identify the shot events. Experimental results have shown our proposed method is effective for basketball event detection.**

***Index Terms*—mean-shift; shot identification; SURF; DBN**

I. INTRODUCTION

In recent years, the basketball game has been more and more popular, the technology on real -time broadcast has developed rapidly and the network has been widely used, all above leads to the remarkable increase of basketball video data which people can get. However it is time consuming and exhausting to watch the whole basketball match video. For common audience, they are only interested in the highlights in the match. But for the expert of the basketball game, they are interested in the video which can reflect the team's tactics or the player's skill. So it is necessary to build an intelligent system which can automatically index and conveniently browse the basketball video data. The key to build such a system is automatic semantic extraction. In this paper we propose a novel basketball processing method that can satisfies this requirement and generate the summaries of the video.

With remarkable development in multimedia systems, many sports applications came into birth. The huge amount of data that is produced by digitizing sports videos demands a process of data filtration and reduction. The large number of sports TV broadcasts also creates a need among sports fans to have ability of seeing interesting parts of all these broadcasts, instead of watching all of them in their entirety. These needs were addressed by the applications such as video summarization and highlight event extraction [1].

Currently, there is some work report in the domain of event identification for soccer video[2], baseball video [4] and tennis video[4]. In [2], Xu eta1 presented a framework for soccer video structure analysis and event detection based on grass-area-ratio. Rui et al. [3] developed effective techniques to detect an excited announcers' speech and baseball hits from noisy audio signals, and fuse them to extract exciting segments of baseball program. In [4], Sudhir et al. presented their techniques on automatic analysis of tennis video to facilitate content-based retrieval, which is based on the generation of an image model for the tennis court-lines.

Semantic analysis of sports video generally involves use of cinematic and object-based features. Cinematic features refer to those that result from common video composition and production rules, such as shot types and replays. Objects are described by their spatial, e. g. , color, texture, and shape, and spatio-temporal features, such as object motions and interactions[5]. Xu *et al.* [6] presents a basketball event detection method by using multiple modalities. Instead of using low-level features, the proposed method is built upon visual and auditory mid-level features, i. e. semantic shot classes and audio keywords, while the computation is tremendous. Nepal *et al.* [7] identified goal segments in a basketball video using five temporal goal models which are constrained by the observation of crowd cheer, scoreboard display and change in direction. However there are some shots that are not meet with the models

In [8], the literature used color and motion information to cluster dominant scene and detect event. In [9], the authors utilize the motion information for describing individual video object, but object segmentation for complex scenes like sports video is still a challenging problem. Thus, we propose an approach to identify the basketball and basket from the complex background.

To detect an d track the ball is the primary work of most of the basketball video analysis. In soccer video, contrasting with the simple background, the ball possession and event detection has been crucially analyzed in [10].

From the above refers and *ref* [11], [12] we can see that most of the research in sports video processing assumes a temporal decomposition of video into its structural units such as clip, scenes, shots and frames

This paper is supported by Natural Science Fund of Shandong (Y2008G09).

doi:10.4304/jsw.6.8.1468-1475

similar to other video domain including television and films. A group of sequential frames often based on single set of fixed or smoothly varying camera parameters (i. e. close-up, medium or long shots, dolly, pan, zoom, etc) form shot. A collection of related shots form scene. A series of related scenes form a sequence. A part of the sequence is called as clip. A video is composed of different story units such as shots, scenes, clips, and sequences arranged according to some logical structure defined by the screen play. In our work, we extract the clips and after analysis assign a descriptive label to each clip and refer the clip as event.

Here we apply SURF (Speed Up Robust Feature)[13] to identify the basketball hoop and use the mean-shift and Kalman filter to identify and track the basketball. Because of the variation of the Shooting environment the objects in the video stream always have many characters such as light-dark variation, rotation, distortion caused by illumination and so on, which make the recognition more difficult. When working with local features, a first issue that needs to be settled is the required level of invariance. Clearly, this depends on the expected geometric and photometric deformations, which in turn are determined by the possible changes in viewing conditions.

The SURF detector is based on the Hessian matrix[8], but uses a very basic approximation, just as DoG (Difference of Gaussians) is a very basic Laplacian-based detector. It relies on integral images to reduce the computation time and we therefore call it the 'Fast-Hessian' detector. The descriptor, on the other hand, describes a distribution of Haar-wavelet responses within the interest point neighborhood.

Based on the basketball and the basketball hoop's low-lever features we may use the DBN(Dynamic Bayesian Net) to find the high-lever semantic events. Our system can identify the shot event that most of the audients are paying close attention to. In the end we will show some experiments results to illustrate the efficiency of our method. The flow chart of the process is as the follow chart: firstly, we segment the video into shots and identify the ball and the hoop in shots using Kalman filter, then the Mean-shift is used to track the ball and the event is identified based on the DBN.

This paper is organized as follows. In section 2 the shot boundary will be detected based on the difference of the frames. In section 3, the SURF features will be extracted for hoop identification. Ball will be identified and tracked using Kalman filter and mean-shift and then the DBN is used to classify the event we defined at section 1. Finally, section 5 concludes the paper and introduces the directions for future research.

II. SHOT BOUNDARY DETECTION

To detect the shot boundary needs to define a variable which value is very big at the boundary while very small in shot. In a shot the adjacent frames' pixel difference is tiny and so is the diagram's standard deviation. If the scene cuts from one shot to another, the difference between them will be very large and so is the standard

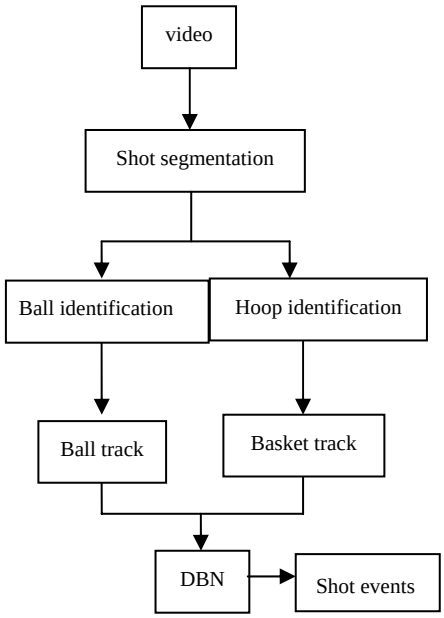


Figure1. The flow chart of the method

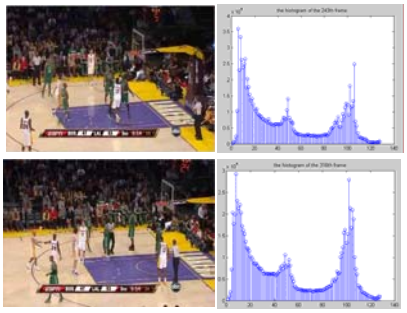
deviation for that reason we select standard deviation to measure the shot transition.

Histogram's standard deviation is defined as the follows:

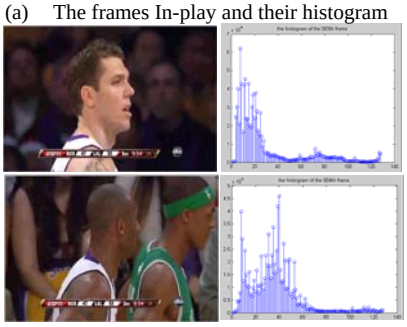
$$\sigma_H(i,i+1)=[\frac{1}{n-1}\sum_{k=1}^{n}(DH_{i,i+1}(k)-\overline{DH})^2]^{\frac{1}{2}} \quad (1)$$

Where $DH_{i,i+1}(k)$ is the difference of the i^{th} and $(i+1)^{th}$ frame's histogram and the $\overline{DH}$ is the average value of the $DH_{i,i+1}(k)$, n is the level of the histogram difference. The first stage used DH value(the difference between the adjacent frames)and a threshold to determine the shot boundary, the threshold is roughly initialized to 0. 4 at the start. Then the luminance of the image is used to eliminate the false positives—when the threshold< L_t (the luminance of the t^{th} frame detected as the boundary) the t^{th} boundary is eliminated.

In Fig. 2 we have shown different types of the frames we classify in our paper and their histograms. (a) are the frames in play and their histogram, (b)are the frames in close up and their histograms, (c) is the difference of the histogram. From the difference of the histogram we can conclude that histogram can be expressed as the feature, at last we draw a curve about the difference of the whole frames histogram by calculating the difference of the adjacent frame. At the boundary the standard is very large as a single peak as shown in Fig. 2.



(a) The frames In-play and their histogram



(b)The frames of close-up and their histogram

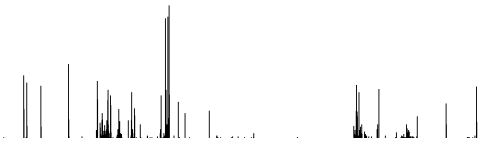


Figure 2. The frames and their histogram of the video and the last figure is the difference between the frames

In the basketball video the low-lever feature include location of the basketball and the position of the basketball hoop, we segment the video into shots and detect the SURF feature of the basketball hoop and track the basketball in one shot. We segment the video stream into shot using the difference between the frames and the luminance feature.

The main purpose of the segmentation is to cut up the video based on the content so as to make it easier to do the follow analysis. The generally method used at the present is based on the difference of features of the adjacent frames, therefore different methods are corresponding to different similarity measurement. Segment the video based on the histogram is widely used because of its compromise the accuracy and the computation, however, the difference of the two frame of which the histogram is very similar to each other though their location is at the shot switch is small and the shot boundary is missed. In response to this situation we propose a method that integrate the information of the luminance of the adjacent frames and extract the differences.

III. FEATURE EXTRACTION FOR HOOP IDENTIFICATION

The basketball hoop tracking is a coarse-to-fine approach. Firstly we choose SURF[6] method to extract features of the backboard and then identify the basketball hoop in the extracted area. Originally the purpose of SURF is used for object recognition, for it is of better speed and accuracy compared with other features, we further extend it to our detection procedure. According to Bay et al. [9], SURF extraction is divided into three steps, which are feature detection, orientation assignment and descriptor components.

A. Feature extraction

Detection of SURF is based on the Hessian matrix, it is defined as follows

$$H = \begin{bmatrix} L_{xx}(\hat{x},\sigma) & L_{xy}(\hat{x},\sigma) \\ L_{xy}(\hat{x},\sigma) & L_{yy}(\hat{x},\sigma) \end{bmatrix} \quad (2)$$

In which $L_{xy} = \frac{\partial}{\partial x^2} g(\sigma) * I(x,y)$, and $g(\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$ Others parameters are similar in the calculation.

Bay *et al.* [13] proposed using box filters to approximate Gaussian second-order derivative and solve the issue of scale space analysis simultaneously. Throughout the processing , the input image is the original image so various scales are not dependent on each other. Additionally, to speed up this algorithm, integral images[14] are used in convolution between box filters and images. The size of box filters directly affects the determinant of the Hessian matrix and scale analysis results.

As an approximation method is used, there must be some impact on the results. It can be corrected as follows [1]:

$$\det(H_{app}) = D_{XX} D_{yy} - (0.9 D_{xy})^2 \quad (3)$$

When generating the image pyramid, there is no need to re-sample the previous results to get a new big scale. Changing scales can be achieved by different of filters. In the algorithm, the scale s satisfies linear changes. In each scale, extreme values are obtained from maxima of the determinant of the Hessian matrix. Then non-maximum suppressions[11] are performed on all candidate extreme points to exclude false-maxima points. Finally, the locations of feature points can be accurately calculated.

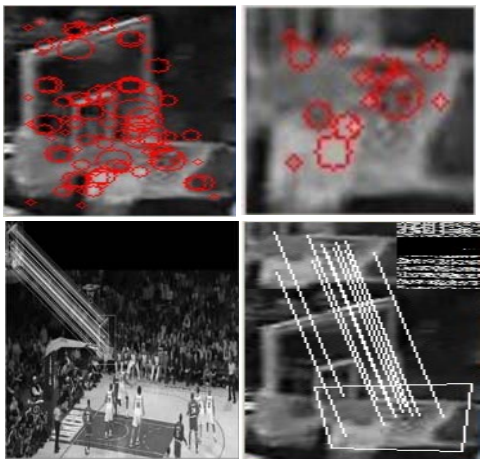
Orientation reflects the direction of dramatic changes of local texture and it can make feature point rotation invariant. An orientation is assigned as follows. Haar wavelet is carried out in the adjacent area of a feature point, each time it accounts for a 60° sector, its result is a 2-D vector, the scope sector rotated continuously until covering the entire circular area, then calculate their 2-norm values of all vectors, choose the maximum one as its orientation.

In order to describe the local texture around a feature point, a 64-D descriptor is established. It stems from a rectangle region around a feature point. This region is further divided into 16 equal parts. In each small region, a 4-D vector is obtained from response of Haar wavelet.

Therefore, the total dimension of the feature descriptor is $4 \times 16 = 64$.

B. Hoop identification

From the video we find that the backboard is entirely different from the background region. In every frame we track the backboard to find the candidate area of the basketball hoop with their matching SURF feature. To find their precise position we apply the SURF feature once again. The tracking result is shown in Fig. 3.



(a) (b)

Figure3. The tracking results

(a) Identify the backboard

(b) Identify the basketball hoop

IV. DETECTION AND TRACKING OF THE BASKETBALL

The aim of an object tracker is to generate the trajectory of an object over time by locating its position in every frame of the video. Object tracker may also provide the complete region in the image that is occupied by the object at every time instant.

A. The ball detection

The tasks of detecting the object and establishing correspondence between the object instances across frames can either be performed separately or jointly. In the first case, possible object regions in every frame are obtained by means of an object detection algorithm, and then the tracker corresponds objects across frames. In the latter case, the object region and correspondence is jointly estimated by iteratively updating object location and region information obtained from previous frames. In either tracking approach, the objects are represented using the shape and/or appearance models.

The model selected to represent object shape limits the type of motion or deformation it can undergo. For example, if an object is represented as a point, then only a translational model can be used. In the case where a geometric shape representation likes an ellipse is used for the object, parametric motion models like affine or projective transformations are appropriate. These representations can approximate the motion of rigid objects in the scene. For a non-rigid object, silhouette or contour is the most descriptive representation and both parametric and nonparametric models can be used to specify their motion.

In this paper Kalman filter is used to build the kinematical model and predict the object's motion so as to narrow the scope of matching, thereby accelerate the speed of target matching. Then the target in the next frame within the specified area is matched in order to establish the association between targets. At last, update the kinematical model to form the moving target chain and get moving target trajectory.

Kalman filter reduces the error covariance matrix in each point in time to the minimum. The procedure includes two parts: (1) prediction, including prediction of the state and prediction of error covariance; (2) modification, including Kalman gain calculation, status modification and error covariance modification. So we can get the fundamental equations of the Kalman filter as follows:

The state prediction equation:

$$\hat{X}_{k|k-1} = A\hat{X}_{k|-1} \qquad (4)$$

Error covariance prediction equation:

$$P_{k|k-1} = AP_{k-1}A^T + Q \qquad (5)$$

Kalman-gain equation:

$$K_k = P_{k|k-1}H_k^T(H_kP_{k|k-1}H_k^T + R_k)^{-1} \qquad (6)$$

State modification equation:

$$\hat{X}_k = \hat{X}_{k|k-1} + K_k(Z_k - H_k\hat{X}_{k|k-1}) \qquad (7)$$

Modification of the covariance equation:

$$P_k = (I - K_kH)P_{k|k-1} \qquad (8)$$

With the Kalman filter we can predict the ball and we can select the filter depending on the actual condition, and then through the above equations estimate the system state.

B. Tracking the ball

Mean Shift is a classic object tracking algorithm with non parameter estimation which is widely used in object tracking. Kalman filter is an algorithm to estimate the linear minimum variance inaccuracy of a dynamic system's states sequence. Its small amount of calculation, real-time computing and starting point's randomness make it very useful in estimation the object's motion. [15] Because of the complexity of the basketball match video, ball detection and tracking is a complex approach. First, Kalman filter is used to predict a position which is to be the initial search center in Mean-Shift algorithm to track. Then, Mean-Shift gets a new target position which is used to be the input parameter of the next Kalman filter.

Tracking the objects by converging at the real location, mean shift continuously iterative calculate the mean-shift vectors by its astringency, for which the tracking algorithm requires the overlap region exist in the target standard and the possible target. In this paper Kalman is used to predict the location of the target to confirm the stability and robustness of the Mean-shift tracking. The

Fig. 3 shows the procedure of the tracking, and the Fig. 4 shows the results of the tracking.

(a) (b) (c)

Fig.4 The identify and track results of the ball

(a) The original frame of the video

(b) Identification of the basketball

(The yellow circle)

(c)Prediction of the basketball

(The red circle)

Mean-shift is a parametric probability density estimation algorithm, which can converge fast at maxima of the probability density function by iteration, therefore it is widely used in the real-time tracking the targets. To the sequence S in n-dimensional Euclidean space X, when $x \in X$ the sample average is defined as the follows: [16]

$$m(x) = \frac{\sum_{s \in S} K(s-x)\omega(s)s}{\sum_{s \in S} K(s-x)\omega(s)} \qquad s \in S \qquad (9)$$

In which K is the kernel function; ω is the sample's weighting function. The difference value $m(x)-x$ is defined as Mean-shift vector. By moving the data point to the Mean-shift vector until it is convergence, the procedure is called Mean-shift algorithm. To the point x in the iteration process, using the kernel function K to calculate the Mean-shift vector's negative grads pointing to the convolution surface $J(x) = \sum_a G(a-x)\omega(a)$, (10)in which the kernel function K and G can satisfy the relation: $g' = -ck(r), r = \| s-x \|^2, c > 0$, g and k are the profile function of G and K, c is a const. After the iteration, the central position of the kernel corresponds to the extreme of a probability density.

V. EVENT ANALYSIS BASED ON THE DBN

Bayesian net is a directed acyclic graph, which reflects the dependencies on the probability inter the variables. Dynamic Bayesian Network combines Bayesian Net with the time information to form a new random model that can deal with time series data based on the Probability Network. It considers factors outside the system as well as the inter-linked inside the system, what's more it dos not only reflect the probability of dependencies between variables but also describes the variation of the variables thus the dynamic model is better than a normal one.

A DBN consist of two parts (a priori Network and Transition Network), having which a DBN will be formed in any length. To facilitate the processing, we suppose the DBN can meet the conditions: (a). the network topology dose not change over time; (b) the network meet the first-order Markov condition. Satisfying the above two conditions the DBN can be seen as unfolding in time sequence.

In the earlier related work, some literature have begun to explore the sports video analysis based on the statistical method, such as the BN (Bayesian Network) and HMM (Hidden Markov Model). Ref [17] used the BN to cluster the soccer video into several typical scenarios categories. In reference [18], an event identification based on the HMM is proposed. However these methods have some limitation on video analysis. Bayesian Net can classify well but can not use the information changing over time, HMM can be applied to the time signal, such as voice signal yet has limitation in model expression for the reason that the video comprise multidimensional signal (temporal and spatial).

Compared with the existing works, we use a stronger temporal signal processor-----DBN (Dynamic Bayesian Network)[19][21]to identify the shot event. On the one hand, considering the transition probability between the various moments, dynamic Bayesian networks extend Bayesian network modeling capabilities of the timing signals. On the other hand, dynamic Bayesian networks allow the use of multiple state variables at the same point in time, while Hidden Markov model uses only one state variable. Based on these considerations, we believe that Dynamic Bayesian Networks is more suitable for sports video content analysis, especially for the semantic analysis of events and their mutual relations.

A. The introduction of the DBN theory

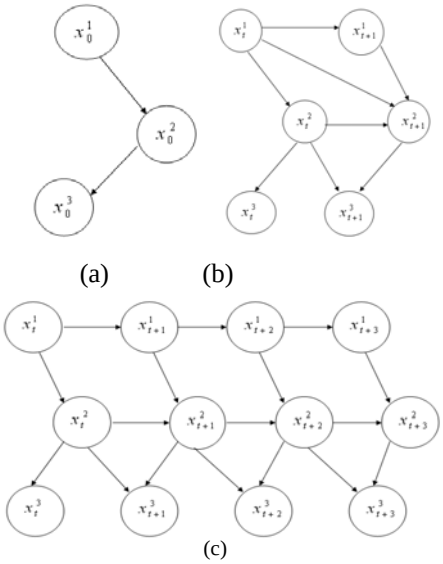
Suppose $P(X_t \mid X_{t-1})$ present the probability of the current status in the condition that the last status has been known. X_t^i is the i^{th} variable's value in t, and $P_a(X_t^i)$ is its parent node. If there are N variables in two status,

$$P(X_t \mid X_{t-1}) = \prod_{i=1}^{N} P(X_t^i \mid Pa(X_t^i)), \qquad (11)$$

Similarly the other nodes' joint distributions are:

$$P(X_{1:T}^{(1:N)}) = \prod_{i=1}^{N} P_{B_0}(X_1^i \mid Pa(X_1^i)) \times \prod_{t=2}^{T}\prod_{i=1}^{N} P_B(X_t^i \mid Pa(X_t^i)) \quad (12)$$

In the context of the basket has been identified, we track the ball's location and measure the Mahalanobis distance[20] between them. When the distance satisfies a condition, we judge the relative position of the previous state and the next state, and then deduce the score event.



(a) (b)

(c)

Fig 5 Schematic diagram of the DBN
(a) The initial network of DBN
(b) The transition network of DBN
(c) Expand the DBN by timeline

B. The training process

The state-space needs some parameters to define its transmission probability model and observation model, and training can estimate the parameters.

In the condition that the network structure is unknown and the samples are totally difference, Friedman proposed the structured EM algorithm[21] The process to learn the probability of the node and the network structure in the condition that the structure and the samples are already known is as follows: (a) add a node into the net; (b) find the optimal network topology according to the node set; repeat the two steps until the network is no longer optimal. Three parameters will be taken into account in the node that is the distance between the ball and net, the pre status of the node, the next status of the node. If the distance at the moment is less than or equal to the threshold, judge the last relative position and the next relative position, if all of them can satisfy the condition, we can count the number of times of the joint appearance of the position to determine the conditional probability as :

$$\begin{aligned} p_t(score) = {} & p_t(score, dis < threshold) \\ & \times p_{t-1}(score, disu < threholdu) \\ & \times p_{t+1}(score, disd < thred) \end{aligned} \qquad (13)$$

C. The event identification

Having the conditional probability and the prior probability for each node, we can interpret the video shots. The video shots are processed by several video analyzers such as the basketball and the basketball hoop identify by the SURF[23] descriptor and the mean-shift algorithm. These low -lever features are extracted as the input of the DBN. We used the evidence propagation procedure from the low-lever features to the relationship features to identify the video shot events and scoring events.

VI. EXPERIMENT RESULTS

In this section, we describe the experiments aimed at evaluating the proposed method, which integrated into a system that was tested using two basketball video sequences. We implement the proposed method on a PC with 2. 8 Pentium D and 1. 5G memory. The material was manually ground-truth and split into clips.

The test videos are two basketball videos NBA games with a total length of fifty minutes, which is segmented into video shots from which we select about 332shots to analyze.

In our experiments results comparison all the ground truths are assigned manually. And we demonstrates the results by using the precision and the recall defined as follows,

$$\text{precision} = \frac{\text{numberof correctdetection}}{\text{numberof correctdetection} + \text{numberof falsedrop}} \qquad (14)$$

$$recall = \frac{\text{number of correct detection}}{\text{number of correct detection} + \text{number of miss}} \qquad (15)$$

A. Shot segmentation

Our proposed shot boundary detection approach has been tested on two matches of basketball from the 2009 All Star NBA matches. The first video is 20 minutes long and contains 127 shots. The second video is of 30 minutes and contains 233 shots. From table 1 the accuracy of our shot boundary detection algorithm is about 93. 7% (the number of correctly detected shots divided by the number of total shots, which is counted manually). The false detection may be caused by photo flash, for instance, if the difference of the histogram of the frames located at the photo flash will be tremendous, consequently, will result in incorrect detection. Table 1 lists the performance in terms of the precision and recall.

TABLE 1. THE RSULT OF SHOTS BOUNDARY DETECTION

	manually	auto	precision	recall
Video 1	147	136	94.4%	93%
Video 2	253	221	96%	92%

The dissimilarity matching on two frames are all based on the histogram of the frames shown in Fig. 2. A promising performance, Recall 85%-93%, and precision 90%-95%, has been achieved.

B. Events detection

After the shot segmentation stage we cluster the shots manually and auto by the method and the results are shown as follows:

In fig. 5 we present a user interface of score events, the system starts after the user's command and uses the parameters get from the input box.

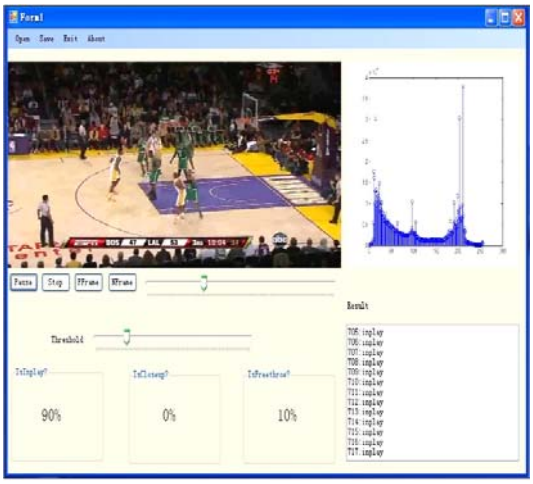


Figure 6 The shot identification system

VII. CONCLUSIONS

We have proposed a novel basketball video shot identification system based on DBN. The main contribution of this paper is to develop an inference system consisting of some linkages between unobservable concepts and observable concepts. We integrate the feature extraction with inferable DBN to fill the gap between the low-level visual domain and the high-level semantic classes. In the experiments, we have proved that our system can identify the shot events and scoring events effectively.

REFERENCES

[1] M. H. Kolekar, K. Palaniappan Semantic Event Detection and Classification in Cricket Video Sequence in Sixth Indian Conference on Computer Vision, Graphics & Image Processing 2008

[2] P. Xu, L. Xie, S-F Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithm and Systems for Segmentation and Strncture Analysis in Soccer Video", In Proc. of IEEE International Conference on Multimedia and Expo, Tokyo, Japan, Aug. 22-25, 2001.

[3] Y. Rui, A. Gupta and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs'', In *Proc. OfACMMultimedia*, Los Angeles, CA, pp. 105-115, 2000.

[4] G. Sudhir, J. C. M. Lee, and A. K. Jain, "Automatic Classification of Tenis Video for High-level Content-based Retrieval", In *Proc.* of *IEEE International workshop on Content-Based Access of Image and Video Database, 1998.* pp. 81-90

[5] Y. Fu, A. Ekin, A. M. Tekalp, and R. Mehrotra, "Temporal segmentation of video objects for hierarchical object-based motion description, " IEEE Trans. Image Processing, vol. 11, pp. 135–145, Feb. 2002.

[6] M. Xu, L. Y. Duan, C. S. Xu, M. S. Kankanhalli, and Q. Tian. "Event Detection in Basketball Video using Multiple Modalities, " Proc. IEEE-PCM'03, Singapore, Dec. 2003.

[7] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of goal segments in basketball videos, "Proceedings of ACM Multimedia Conf. on Authoring Support, Oct. 2001.)

[8] H. Lu and Y. -P. Tan, "Content-based sports video analysis and modeling, " in Proceedings of 7th International Conference on Control, Automation, Robotics and Vision (ICARCV '02), pp. 1198–1203, Singapore, December 2002.

[9] Y. Fu, A. Ekin, A. M. Tekalp, and R. Mehrotra, "Temporal segmentation of video objects for hierarchical object-based motion description, " IEEE Transactions on Image Processing, vol. 11, no. 2, pp. 135–145, 2002.

[10] X. Yu, Q. Tian, and K. W. Wan. "A novel ball detection framework for real soccer video, " Proc. IEEE-ICME'03, pp. 265-268, 2003.

[11] V. Chasanis, A. Likas, and N. Galatsanos. Scene detection in videos using shot clustering and symbolic sequence segmentation. *in IEEE Workshop on Multimedia Signal Processing*, 2007.

[12] C. Ngo, T. Pong, and H. Zhang. On clustering and retrieval of video shots through temporal slices analysis. *in IEEE Transactions on Multimedia*, 4(4), 2002.

[13] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. Proceedings of Ninth European Conference on Computer Vision(ECCV), Graz, May 7, vol. 13: pp. 404-417, 2006.

[14] Mikolajczyk, K. , Schmid, C. : Indexing based on scale invariant interest points. In: ICCV. Volume 1. (2001) 525 – 531)

[15] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. Proceedings of Ninth European Conference on Computer Vision(ECCV), Graz, May 7, vol. 13: pp. 404-417, 2006.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features, " Conference on Computer Vision and Pattern Recognition, pp. 511-518, 2001.

[17] David G. Lowe, "Distinctive Image Features from Scale-Invariant Key points, " J, International Journal of Computer Vision (IJCV), vol. 60, pp. 91-110, 2004.

[18] IU Ji-yan, PAN Jian-shou, WU Ya-peng, et al. Mean-Shift tracking algorithm combined with Kalman filter. Computer Engineering and Applications, 2009, 45（12）: 184-186. 2. 3 Detection of the player

[19] Cheng Y. Mean shift, mode seeking and clustering[J] IEEE Transactions on Pattern Analysis and Machine Intelligence. 1995. 17(8): 79 — 799

[20] M. Luo, Y. F. Ma, and H. J. Zhang, "Pyramidwise Structuring for Soccer Highlight Extraction, "Proceedings of IEEE Pacific-Rim Conference on Multimedia, 2003.

[21] G. Xu, Y. F. Ma, H. J. Zhang, and S. Q. Yang, "A HMM Based Semantic Analysis Framework for Sports Game Event Detection, "Proceedings of IEEE International Conference on Image Processing, 2003.

[22] K. P. Murphy, "Dynamic Bayesian Network Representation, Inference and Learning, " PhD Dissertation, University of California, Berkeley, 2002.

[23] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2001). "8. 5 The EM algorithm". The Elements of Statistical Learning. New York: Springer. pp. 236–243. ISBN 0-387-95284-5.

Yun Liu received the B. E. degree in automatic control from Taiyuan Science & Technology University, Shanxi, P. R. China, in 1982, the M. S. degree in electric engineering from Tongji University, Shanghai, China, in 1995 and the Ph. D degree in signal and information process from China Mining & Technology University, Beijing, China, in 2001. His main research interests are digital signal processing, digital image processing and multiple media communication.

XueYing Liu was born on June 26, 1984, in Shandong Province, China. She is a postgraduate student of Qingdao University of Science and Technology. Her research interests are digital signal processing and digital image processing.

Chao Huang was born on Novemeber 12, 1981, in Jiangxi Province, China. He is a postgraduate student of Qingdao University of Science and Technology. His research interests are digital signal process and computor software.