Baseball and Tennis Video Annotation with Temporal Structure Decomposition

Jui-Hsin Lai [1] and Shao-Yi Chien [2]

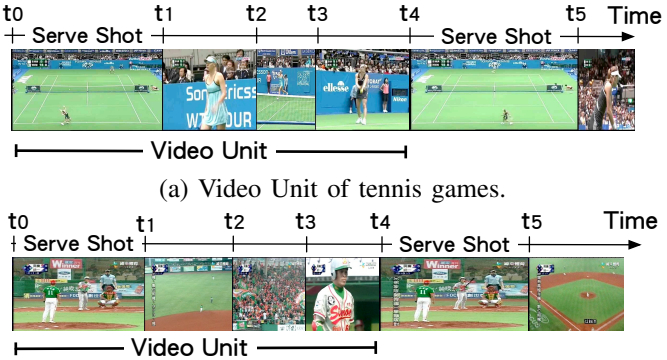
Media IC and System Lab
Graduate Institute of Electronics Engineering and Department of Electrical Engineering
National Taiwan University
BL-421, 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan
[1] larry@media.ee.ntu.edu.tw, [2] sychien@cc.ee.ntu.edu.tw

***Abstract*—Sport video annotation can help viewers easily browse sport video content and quickly find the hot events and highlights in a game. Although many annotation algorithms have been proposed, they are not suitable for practical implementation since the high complexity and the low precision rates are not acceptable. In this paper, a method of sport video temporal structure decomposition, which decomposes the sport video into many video clips, is proposed. Then score box information and additional semantic information are important clues for event annotation. Experimental results show that the proposed algorithm can successfully and effectively decompose video into clips. The annotation results also have extremely high precision and recall rates for both baseball and tennis videos.**

I. Introduction

Facing the large quantity of sport videos, viewers can easily browse the game and find out hot events and highlights with video analysis tools. Several previous works on sport video annotation can be found in literatures. For algorithms with Hidden Markov Model (HMM), Chang et al. employ HMM to extract highlights from baseball video [1], and Kijak et al. parse tennis video structure with HMM [2]. However, the precision and recall rates are still not high enough, and the HMM model needs large training data sets to construct the model and requires complex computation. Lie and Shia combine captions and visual features for semantic event detection of baseball video [3]. Nevertheless, that needs a framework to analysis the video structure. The detected events are hard to be extracted without the analysis of video structure.

In our observation, the game proceedings of some sport videos, like baseball, badminton, table tennis, tennis, and volleyball, are repeating the iteration : player serve, game running, and game stop. We call these sport videos as serve-running-stop sport videos. The serve-running-stop sport videos are well structured and can be decomposed into many video clips. One serve-running-stop video clip can be seem as a Video Unit, as shown in Fig.1. Each Video Unit usually presents an event in sport video. It begins at a serve shot, such as a pitch shot in baseball or a rally shot in tennis, and ends before the next serve shot. With such a regular structure, the annotation process can be done by decomposing the whole video into Video Units on temporal domain and then extracting semantic information from each Video Unit. So, finding the serve shots is the key step for video decomposition.



(a) Video Unit of tennis games.

(b) Video Unit of baseball games.

Fig. 1. Video Unit in sport video.

We propose an algorithm to recognize each serve shot in video using query-by-example. The method of query-by-example employs histogram of serve shot as the example and uses this example to query the whole video looking for the shots with histogram similarity. To annotate the sport event of each Video Unit, score box recognition and semantic information extraction are applied for event judgment. For experimental results, the temporal decomposition with query-by-example can achieve high recognition rates, decrease large complex computation, and be suitable for all serve-running-stop sport videos. The annotation results of baseball and tennis videos also have extremely higher precision and recall rates than previous works.

II. Proposed Annotation Scheme

The proposed annotation scheme, as shown in Fig. 2. First of all, shot boundary detection detects the scene change boundaries and records the time stamps, which are shown as $t_0, t_1, t_2, ...$ in Fig. 1. According to the time stamps, serve shot recognition employs the query-by-example to detect the serve shot. After that sport videos are temporally decomposed into Video Units. Serve shot update will dynamically compensate histogram shift due to the change of camera diaphragm and

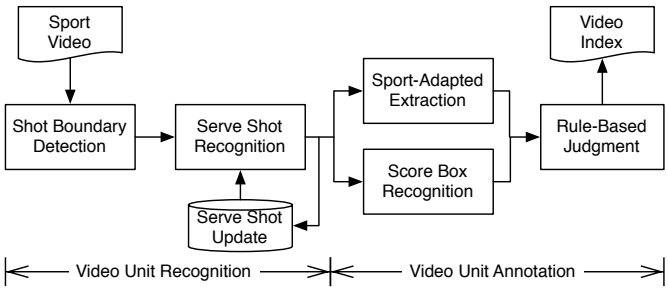


Fig. 2. Proposed sport video annotation scheme.

update the query example following the game proceeding, which can maintain the high serve shot recognition rate.

Score box recognition recognizes and records the score box information from every Video Unit. To achieve higher level semantic annotation, sport-adapted extraction is required to get field, player, and ball information. Rule-based judgment then combines the above information and judges the event happened in a Video Unit. Finally, the annotation scheme outputs the sport video index to viewers.

III. Recognition of Video Unit

In our observation, the Video Unit is the basic component of a serve-running-stop sport video and usually presents a sport event. Every Video Unit begins with a serve shot, such as pitch shot in baseball games or rally shot in tennis games, so finding the serve shot is the key operation to segment a video into several Video Units.

A. Shot Boundary Detection

To detect the serve shots, shot boundary detection is the first step. Some previous works have been proposed to detect the abrupt and dissolve scene changes. For sport videos, the serve shots usually have steady camera motion. Based on this observation, we can simplify the detection methods [4] and employ the difference of luminance histogram to detect the shot boundary. Suppose that the luminance distributions are discretized into m-bins. The operation of histogram difference $\mathbf{H}$ between adjacent frames can be calculated in the following equation.

$$\mathrm{H} = \sum_{i=1}^{m} \|Y_f(i) - Y_{f-n}(i)\|, \tag{1}$$

where Y_f is the histogram of frame index f, and Y_{f-n} is the histogram of frame index $f-n$. Note that, the adjacent frames are continuous frames with an interval of n, where n = 1, 2, ..., N. To detect the dissolved scene change, the larger n is applied to increase histogram difference $\mathbf{H}$. Then, a threshold value is set to detect scene change. For threshold setting, the lower threshold value may make error boundary detections at camera in fast motion, but these errors do not make the error recognition in the following serve shot recognition.

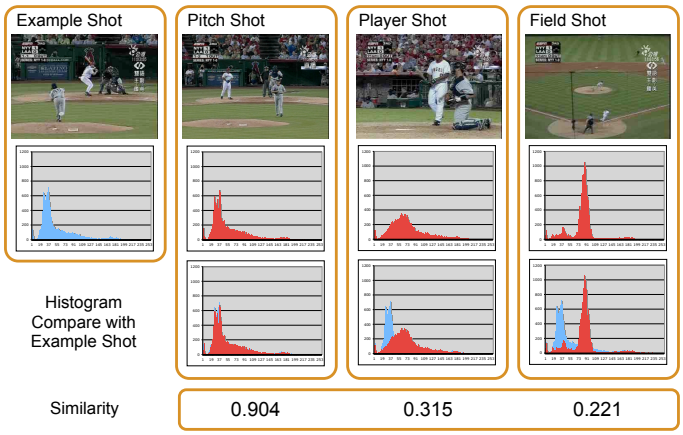


Fig. 3. Serve shot recognition with similarity calculation between each shot histogram and the example shot histogram.

B. Serve Shot Recognition

To recognize the serve shot, the previous work uses camera motion, image edge distribution, edge intensity and color distribution to be the clues [5]. However, the computation of clue extraction is complex and needs multiple thresholds, and the recognition rate is not high enough. In sport video, the serve shots usually have the unique histogram distribution. Based on this observation, serve shot recognition employs one serve shot histogram as the example to query the video looking for shots with histogram similarity.

There are some popular methods of similarity measure like the Bhattacharyya coefficient [6] and the earth mover's distance (EMD) [7]. Nevertheless, the threshold setting of Bhattacharyya coefficient is case-sensitive and needs manual adjustment for different sport video. The EMD method takes complex computation and needs lots recursive loops to generate the results. In this article, the region match of histogram bin height is employed. Suppose that the histogram of luminance distribution is discretized into m-bins, and the histogram of color distribution is discretized into n-bins. The similarity $S[p,q]$ is calculated between the current shot histogram $p(i)$ and the example histogram $q(i)$, as shown in the following equation.

$$\mathrm{S}[p,q] = \frac{\#\{p(i) \mid q(i)/k < p(i) < kq(i), \forall i \in [1, m+n]\}}{m+n}, \tag{2}$$

where variable k is the scale factor of region match, and the match index i is for both luminance and color histograms. The larger $S[p,q]$ is, the more similar the distributions are. For two identical histogram distributions, $S[p,q]$ is equal to 1. Figure 3 illustrates an example of serve shot recognition, where pitch shot, player shot, and field shot are queried by the example shot. It shows that each shot category has an unique histogram distribution, and only the pitch shot can obviously have higher similarity. For the initialization of example histogram assignment, manually storing one serve shot as the example histogram is used.

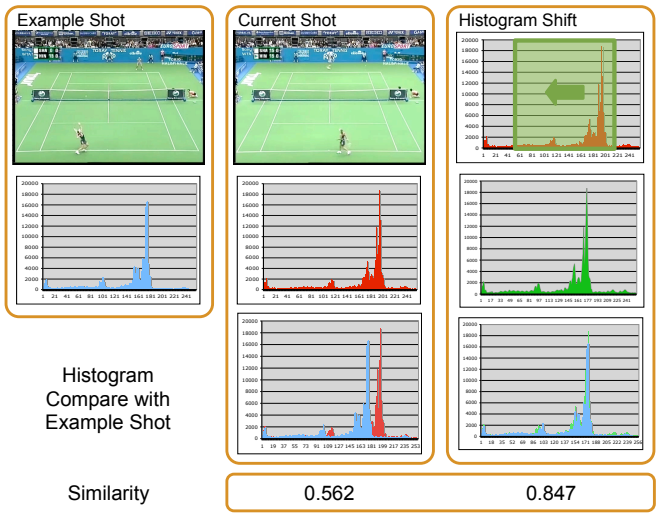


Fig. 4. Histogram shift compensation for camera diaphragm change.

C. Histogram Shift Compensation and Serve Shot Update

Sometimes, with camera zooming in and zooming out, the camera diaphragm would change and make the luminance histogram shifted by a small distance. The histogram shift can be modeled as a linear shift for slight luminance change. For each shot, the luminance change is compensated by shifting the histogram with a distance D at the similarity calculation. The distance D is the difference of the luminance bin indexes with maximum height of the example shot $q(i)$ and the current shot $p(i)$, which can be shown as the following equation.

$$D = \arg \max_{1 \le i \le m} p(i) - \arg \max_{1 \le i \le m} q(i) \quad (3)$$

In Fig. 4, the similarity is largely increased after histogram compensation, which can improve the precision of the proposed algorithm. By the game going on, the histogram distributions of a scene would slowly change. Thus, the update of example shot histogram is implemented by the equation

$$q_t(i) = (1-\alpha)q_{t-1}(i) + \alpha p_{t-1}(i), \forall i \in [1, m+n], \quad (4)$$

where α weights the contribution of current serve shot. The update equation evokes a forgetting process in the sense that the contribution of a specific shot decreases exponentially the further it lies in the past. Taking advantage of serve shot update, the recognition process is effective from the beginning to the end of a game.

IV. Annotation of Video Unit

The score box in sport video gives viewers a lot of information, and therefore this is an important clue for event annotation. Combining more semantic information, the event of each Video Unit is annotated.

A. Score Box Recognition

The score box usually has the same style, character type, and exists in the corner of screen in a broadcasting. Therefore, we can set score box location and save all number and character images in advance as a prior knowledge. To recognize the numbers in score box, template matching is employed to find the best matched pre-stored number images. With this method, the numbers of ball, strike, out, and base occupation in baseball games and the score numbers and server in tennis games are extracted.

B. Sport-Adapted Information Extraction

There are still some sport events cannot be distinguished only by score box information, like an event is hit or walk in baseball games, and an event is ace or normal rally in tennis games. To distinguish these events, some additional feature extractions are required. In baseball video, the camera usually takes the outfield scene when hit and field out events happened, and it must have green field shots in the Video Unit. So, the green field detection is the additional feature in baseball videos. In tennis video, the higher semantic information, like ball and players trajectories, would be extracted by our previous work [8].

C. Rule-Based Judgment

The rule-based judgment combines the above information to detect tennis events: break point, double fault, volley, rally, and ace. For example, an ace event is judged as server get score, and the ball trajectory cross net one time. A volley event is judged as ball hitting near the net. In baseball games, hit, walk, strike out, field out, stolen base, double play, 2B, 3B, and home run can also be judged. For instance, a hit event can be defined as green field scene detection and base occupation increase or score increase in the out number fixedly.

V. Experimental Results

Different sport videos with two hours of Chinese Professional Baseball League (CPBL), three hours of Major League Baseball (MLB), two hours tennis game of Australia Open women's single (AOWS), and three hours tennis game of Australia Open men's single (AOMS) are used as test patterns. Before the start of annotation, we need to manually store one serve shot to be the example shot, save number/character images for score box template matching, and set the position of the score box in advance as a prior knowledge.

The experimental results of serve shot recognition, with the scale factor k = 1.25 and update rate α = 0.2, are shown in Fig. 5, which are the figures of correlations between similarity thresholds and recognition rates. For Fig. 5(a)(c)(e), we can see that the higher precision rates appear at higher similarity threshold, and the higher recall rates appear at lower similarity threshold. That is hard to get both higher recall and precision rates at the same similarity threshold for all sport videos. However, with the help of serve shot update, the recall rates can be increased under higher similarity threshold in Fig. 5(b)(d)(f). It is easy to set a similarity threshold to get both high precision and recall rates, and it also shows the proposed methods are effective for all the testing sequences. For processing time of serve shot recognition, the proposed methods just need six minutes to process one hour video on

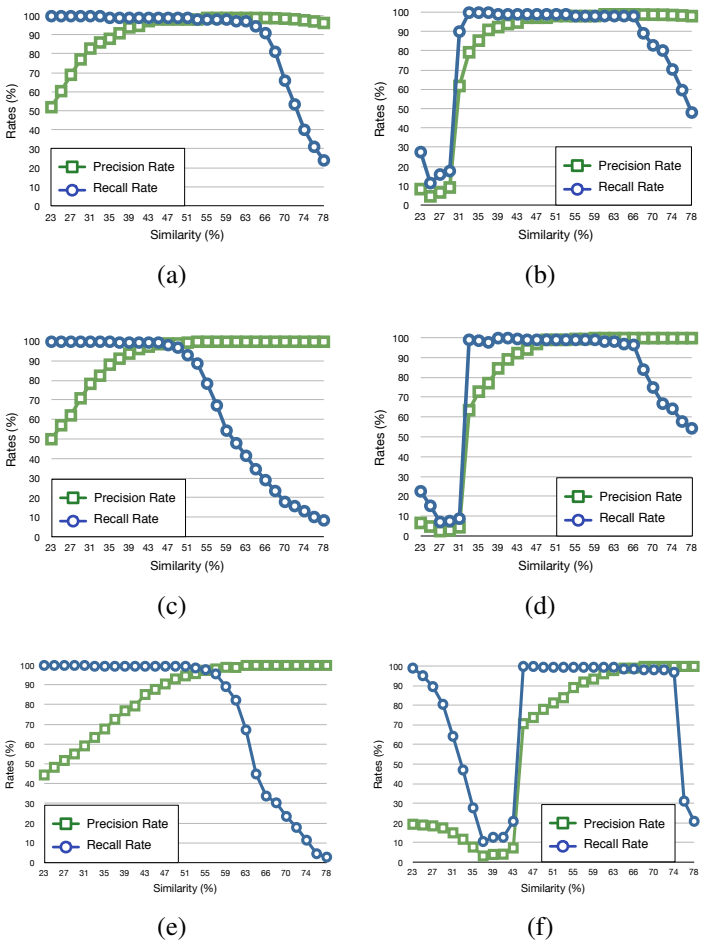


Fig. 5. The recall and precision rates of serve shot recognition under different similarity thresholds. (a) The rates without serve shot update of AOWS. (b) The rates with serve shot update of AOWS. (c) The rates without serve shot update of AOMS. (d) The rates with serve shot update of AOMS. (e) The rates without serve shot update of MLB. (f) The rates with serve shot update of MLB.

TABLE I
RESULTS OF BASEBALL EVENT ANNOTATION.

Event	Quantity	Recall (%)	Precision (%)
Home Run	1	100	100
Stolen Base	2	100	100
Hit	24	87.5	100
Walk	11	100	100
Strike Out	11	72.7	88.9
Field Out	50	96	92.3
Total	99	91.9	94.8

a PC with Pentium IV 3 GHz CPU, which is ten times faster than real-time requirement.

For Video Units annotation, score box recognition and sport-adpated information extraction can provide high confidence information to be the clues. Relying on high recall and precision rates in serve shot recognition, the annotation result can maintain high recall and precision rates. The results

TABLE II
RESULTS OF TENNIS EVENT ANNOTATION.

Event	Quantity	Recall (%)	Precision (%)
Break Point	11	100	100
Double Fault	3	100	100
Volley	8	100	100
Rally	43	97.7	91.3
Ace	14	71.4	76.9
Total	79	93.7	91.4

of baseball video annotation are shown in Table I, where the average recall rate is 91.9% and precision rate is 94.8%. The results of tennis video annotation are also shown in Table II, where the average recall rate is 93.7% and precision rate is 91.4%. The whole experimental results are all better than previous works [1][2][3].

VI. CONCLUSIONS

In this paper, we propose an annotation scheme suitable for serve-running-stop sport video and apply this scheme on baseball and tennis videos as experiments. The sport video is decomposed into Video Units, and then each Video Unit is annotated. For video temporal decomposition, query-by-example achieves high precision rate and recall rate of serve shot recognition with low complex computation. In addition, score box information and additional semantic information are important clues for event judgment. Finally, it was shown that rule-based judgment can achieve very high precision and recall rates of event annotation.

REFERENCES

[1] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden markov models," in *Proceedings of the International Conference on Image Processing 2002*, vol. 1, Sept. 2002, pp. I–609 – I–612.

[2] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot, "Hmm based structuring of tennis videos using visual and audio cues," in *Proceedings of the International Conference on Multimedia and Expo 2003*, vol. 3, July 2003, pp. III–309–312.

[3] W.-N. Lie and S.-H. Shia, "Combining caption and visual features for semantic event classification of baseball video," in *Proceedings of the International Conference on Multimedia and Expo 2005*, July 2005, pp. 1254 – 1257.

[4] W. A. C. Fernando, C. N. Canagarajah, and D. R. Bull, "Scene change detection algorithms for content-based video indexing and retrival," *Electronics and Communication Engineering Journal*, vol. 13, no. 3, pp. 117–126, June 2001.

[5] J. Assfalg, M. Bertini, C. Colombo, and A. D. Bimbo, "Semantic annotation of sport video," *IEEE Multimedia*, vol. 9, no. 2, pp. 52–60, April-June 2002.

[6] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *IEEE Transactions on Communications Technology CMO-15(1)*, vol. 15, no. 1, pp. 52–60, Feb. 1967.

[7] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 840–853, May 2007.

[8] J.-H. Lai and S.-Y. Chien, "Video program: Tennis video 2.0: a new framework of sport video applications," in *Proceedings of the 15th International Conference on Multimedia MULTIMEDIA '07*, September 2007, pp. 1087–1088.