NOVEL FRAMEWORK FOR SPORTS VIDEO ANALYSIS: A BASKETBALL CASE STUDY

Chun-Min Chen and Ling-Hwei Chen

Department of Computer Science, National Chiao Tung University
Hsinchu, Taiwan, R.O.C.
E-MAIL: cmchen.nctu@gmail.com, lhchen@cc.nctu.edu.tw

ABSTRACT

Semantic event and slow motion replay extraction for sports videos have become hot research topics. Most researches analyze every video frame; however, semantic events only appear in frames with scoreboard, whereas replays only appear in frames without scoreboard. Extracting events and replays from unrelated frames causes defects and leads to degradation of performance. In this paper, a novel framework is proposed to tackle challenges of basketball video analysis. In the framework, a scoreboard detector is first provided to divide video frames to two classes, with/without scoreboard. Then, a semantic event extractor is presented to extract semantic events from frames with scoreboard and a slow motion replay extractor is proposed to extract replays from frames without scoreboard. Experimental results show that the proposed framework is practicable for basketball videos. It is expected that the proposed framework can be extended to other sports.

***Index Terms*—Basketball, broadcast video, semantic event extraction, slow motion replay detection, sports video analysis**

1. INTRODUCTION

Thanks to the rapid growth of computer science and network technology, people now are capable of using mobile devices, e.g. notebook, tablet, smart phone, to acquire sports videos anytime and anywhere. However, substantial number of sports videos are produced and broadcasted every day. It is nearly impossible to watch them all. Most of the time, people prefer to watch highlights of sports videos or retrieve only partial video segments that they are interested in. Therefore, sports video analysis, such as semantic event extraction [1]-[9] and slow motion replay detection [10]-[18], has become a valuable and hot research topic.

A. Semantic Event Extraction Challenges

Some semantic event extraction researches [1]-[3] use video content as resource knowledge. However, schemes relying on video content encounter a challenge called semantic gap, which represents the distance between video features and semantic events. Recently, some researches [4]-[9] use a multimodal fusion of video content and external resource knowledge to bridge the semantic gap. Webcast text, one of the most powerful external resource knowledge, is an online commentary posted with well-defined structure by professional announcers. It focuses on sports games and contains detail information (e.g., event description, game clock, player involved, etc.). The multimodal fusion scheme, which analyzes webcast text and video content separately and then does text/video alignment to complete sports video annotation or summarization, has been used in American football [4], soccer [6]-[8], and basketball [7]-[8].

In the multimodal fusion scheme, text/video alignment can be performed through video game clock recognition. Xu et al. [6]-[8] used Temporal Neighboring Pattern Similarity (TNPS) measure to locate game clock and recognize each digit of the clock. A detection-verification-redetection mechanism is proposed to solve the problem of temporal disappearing clock region in basketball videos. However, recognizing game clock in a frame which has no game clock is definitely unnecessary. The cost of verification and redetection could have been avoided. Moreover, the clock digit characters cannot be located on a semi-transparent scoreboard.

B. Slow Motion Replay Detection Challenges

For slow motion replay detection, many methods have been proposed, and they can be classified into two categories. The first category [10]-[15] is to locate positions of specific production actions called special digital video effects (SDVEs) or logo transitions, and bases on these positions to detect replay segments. However, in this category, they all made an imperfect assumption that a replay is sandwiched by either two visually similar SDVEs or logo transitions, the assumption is not always true in basketball videos. In fact, a basketball video segment bounded by paired SDVEs is not always a replay. Moreover, the beginning and end of a basketball replay can have some combinations: 1) paired visually similar SDVEs; 2) non-paired SDVEs; 3) a SDVE in one end and an abrupt transition in the other. So, previous work in this category cannot be applied to basketball videos with replays having combinations (2) and (3).

The second category [16]-[18] analyzes features of replays to distinguish replay segments from non-replay segments. Farn et al. [16] extracted slow motion replays by referring to the dominate color of soccer field; however, it is not applicable in basketball videos since the size of basketball court is relatively smaller and its textures are more complicated. Wang et al. [17] conducted motion-related features and presented a support vector machine (SVM) to classify slow motion replays and normal shots. The precision rates of two experimented basketball videos are 55.6% and 53.3% with recall rates 62.5% and 66.7%, respectively. Han et al. [18] proposed a general framework based on Bayesian network to make full use of multiple clues, including shot structure, gradual transition pattern, slow motion, and sports scene. The method is suffered from the inaccuracy of the used automatic gradual transition detector. Their experiments show precision rate 82.9% and recall rate 83.2%.

The existing two category methods are generic but not satisfactory for basketball videos. Moreover, most previous researches analyze every video frame to detect replays, but

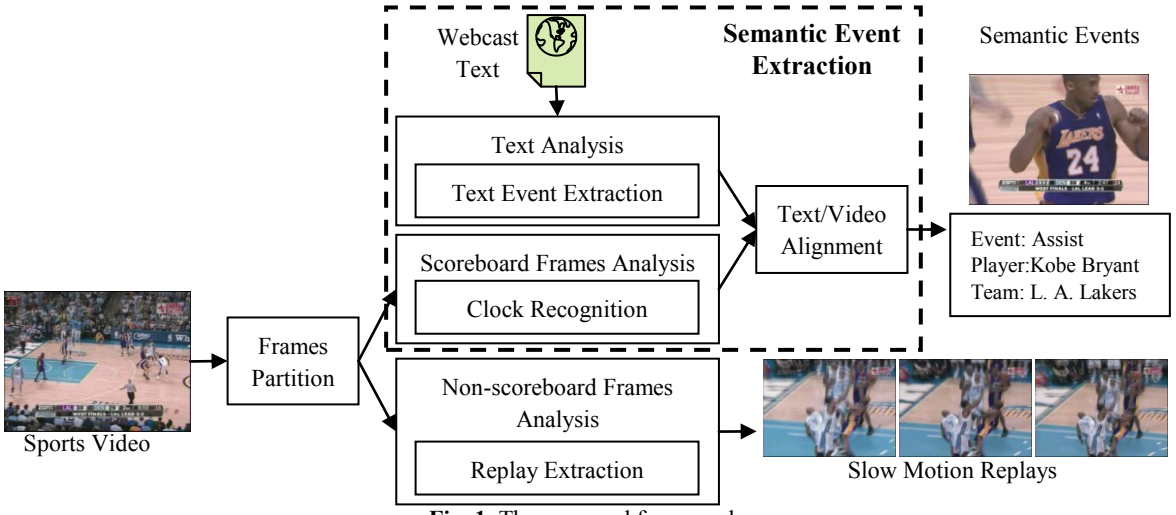


Fig. 1. The proposed framework.

detecting replays in video frames that are surely non-replay degrades both performance and detection rate.

To tackle above-mentioned challenges of sports video analysis, in this paper, we propose a novel framework to analyze basketball videos. One of the main novelties is to refer to scoreboard information. It is observed that sports video frames can be partitioned into two categories according to the existence of scoreboard. Frames with scoreboard existence are called scoreboard frames, and others are called non-scoreboard frames. In general, semantic events appear during playing of a sports game, which consists of scoreboard frames only. Slow motion replays appear during temporal pausing of a sports game, which consists of non-scoreboard frames only. The phenomenon is dominant and used to skip large amount of unnecessary processing frames before semantic resource extraction. Accordingly, the performance and the detection rate can be assured.

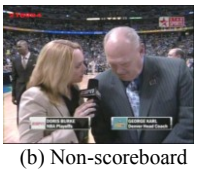
The proposed framework shown in Fig. 1 is not only capable of acquiring both two valuable semantic resources in one time, but also increases detection rate for semantic resources extraction. In Section 2, a video frame partition method is introduced to divide frames into scoreboard frames and non-scoreboard frames. Extracting semantic events from scoreboard frames and extracting slow motion replays from non-scoreboard frames are presented in Section 3 and Section 4 respectively. Experiments and conclusions are given in Section 5 and Section 6.

2. VIDEO FRAMES PARTITION

As can be seen from Fig. 2, in basketball videos, all frames can be broadly classified into two categories, scoreboard frames and non-scoreboard frames. Scoreboard frames present basketball game with scoreboard overlaid on them, while non- scoreboard frames present the rest, e.g., sideline interview, slow motion replay, etc. Since semantic events only appear in scoreboard frames, whereas replays only appear in non-scoreboard frames. It is beneficial to filter out unnecessary processing frames in each semantic resource extraction step. So, an automatic scoreboard template extractor is needed. As shown in Fig. 2(a), a scoreboard is fixed rectangular area with pixels changing infrequently. Based on this fact, our previous work [19] presented an automatic scoreboard template extractor. Here, we adapt this extractor to get the scoreboard template and position. After scoreboard template extraction, the video frames partition can be done by matching every frame with scoreboard template at the scoreboard position.

Contrary to previous works, in the paper, scoreboard frames and non-scoreboard frames will be separately processed in semantic event extraction and slow motion replay detection. Since scoreboard only covers a small part of a video frame, conducting this slight-cost partitioning task before semantic resource extraction improves a lot of performance in both time complexity and detection accuracy.

(a) Scoreboard frame.



(b) Non-scoreboard frame due to sideline interview.

(c) Non-scoreboard frame due to player's statistic.

(d) Non-scoreboard frame due to slow motion replay.

Fig. 2. Examples of scoreboard frames and non-scoreboard frames.

3. SEMANTIC EVENTS EXTRACTION FROM SCOREBOARD FRAMES

It can be seen from Fig. 1, a multimodal fusion scheme is conducted for semantic events extraction from scoreboard frames. Using our previous work [20], text events with game clock can be extracted from webcast text. Then, a text/video alignment and event annotation method is proposed by recognizing game clocks of scoreboard frames.

As to game clock recognition, location of each clock digit is first located. A digit templates collection scheme is provided to collect digit character templates. With clock digit locations and

digit templates, a two-step strategy is proposed to recognize game clocks on the semi-transparent scoreboard in scoreboard frames. Here, without loss of generality, four game clock patterns can be defined in Fig. 3.

Game Clock Patterns	X_1X_2:X_3X_4	X_2:X_3X_4	X_3X_4.X_5	X_4.X_5
Meaning of Each Digit	X_1:TEN-MINUTE, X_2:MINUTE, X_3:TEN-SECOND, X_4:SECOND, X_5:TENTH-SECOND			

Fig. 3. General definitions of game clock patterns.

3.1. Clock Digit Locator

Based on the fact that there are 30 frames and 300 frames taken in one second and ten seconds, for each scoreboard frame f_i, the pixel-based frame difference between f_i and f_{i-30} and that between f_i and f_{i-300} are first calculated as follows:

$$Df_{i,30}(x,y) = \left| f_i(x,y) - f_{i-30}(x,y) \right|. \quad (1)$$

$$Df_{i,300}(x,y) = \left| f_i(x,y) - f_{i-300}(x,y) \right|. \quad (2)$$

Where (x, y) is a point of the scoreboard area. Then, two accumulated difference frame, $ADf_{i,30}$ and $ADf_{i,300}$, are created by

$$ADf_{i,30}(x,y) = \sum_{j=31}^{i} Df_{j,30}(x,y). \quad (3)$$

$$ADf_{i,300}(x,y) = \sum_{j=301}^{i} Df_{j,300}(x,y). \quad (4)$$

The accumulated difference at each pixel can be considered as the change degree at that position. Since SECOND digit changes every 30 frames and TEN-SECOND digit changes every 300 frames, two approximated areas of SECOND digit and TEN-SECOND digit can be located by observing $ADf_{i,30}$ and $ADf_{i,300}$ in scoreboard frame sequences. Based on positions and sizes of these two areas, a complete game clock area is located.

Note that each game clock pattern consists of a separation mark (colon or dot). It is observed from the vertical projection histogram of the game clock area that the separation mark is located at the lowest local peak, and all clock digits are separated from each other by a local valley. Based on the information and the width of the detected SECOND digit area, each clock digit area can be located. An example is shown in Fig. 4.

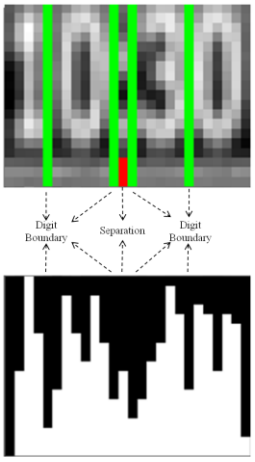


Fig. 4. An example of locating game clock digits (10:30).

3.2. Clock Digit Template Collection

After locating each clock digit area, a simple idea is proposed to collect clock digit templates. Every time TEN-SECOND digit changes, it means SECOND digit area has been through a complete cycle from 9 to 0, i.e. 9, 8, 7, …, 0. Therefore, once a pattern change of TEN-SECOND digit area is detected, a set of digit templates can be collected by sampling from SECOND digit area every second. After collecting eleven samples in consecutive eleven seconds as a candidate template set, a verification step is provided to examine whether the set is correct. For each candidate template set, if the first template is exactly the same as the eleventh one, then the candidate is considered as a correct digit templates set since all members in the set complete a cycle. Otherwise, keep collecting another candidate set until a correct one is verified.

3.3. Clock Digit Recognition

After locating each clock digit and collecting a correct digit templates set, a two-step strategy is proposed to recognize game clocks on the semi-transparent scoreboard.

First, for each clock digit, a local strategy is proposed to narrow the number of digit templates used in pattern matching. While applying template matching to recognize each digit, only the following three candidates should be considered: 1) current digit character; 2) next digit character; 3) possible digit character derived from frame number difference. For example, if we try to update SECOND digit of "10:30", the first candidate is "0" itself. The second candidate is "9". Assume it has been 60 frames since the last recognized result of SECOND digit, and frame rate of video is 30. It is possible that two seconds has been collapsed since the last update, so the third candidate is derived to be "8". Note that narrowing the number of candidates not only prevents possible errors from matching other digit templates, but also provides a mechanism to correct the recognition result in later frames.

After applying local strategy in pattern matching for each clock digit, a global strategy is proposed to verify the overall game clock recognition result. For example, if we recognize the clock time as *mn:st*, the overall recognition result is

$$T = (m \times 10 + n) \times 60 + s \times 10 + t.$$

For each recognition result of frame f_k, $T(k)$, and a new candidate result recognized from a later frame f_l, $T(l)$, the verification equation is defined by

$$T(k) = T(l) + Round((k - l) \div 30).$$

If the equation holds, the new candidate result is regarded as a right one, and the frame f_l is recognized as game clock $T(l)$. Note that the application here focuses on semantic event extraction, so recognition result for every single frame is not important. Instead, recognizing when a right game clock is updated is more valuable for text/video alignment and event annotation.

3.4. Text/Video Alignment

Based on the recognized game clock, a text/video alignment is presented to do sports video annotation. The alignment consists two parts. First, through the recognized game clock in video frames, the corresponding target frame of each event extracted from webcast text [20] is located, this is called event moment detection. Second, the time period for each event is determined, this is called event boundary detection.

As to event boundary detection, here, we set a general interval for all kinds of basketball events. For example, ten seconds before the event moment to five seconds after the event moment.

4. REPLAY DETECTION FROM NON-SCOREBOARD FRAMES

After partitioning video frames, every consecutive non-scoreboard frame sequence can be considered as a non-scoreboard segment (NS). According to our previous work [19], there are only two different kinds of segments for non-scoreboard segments: 1) slow motion replay; and 2) game-related segment which shows game-related information with background around the court but is not a replay. Some game-related segment examples are given in Fig. 2(b)-2(c). Here, characteristics of replays and non-replays are observed to create features, which are used in the k-means cluster with k=2 to detect replays and prune non-replays.

Since a slow motion replay is never less than 6 seconds, it is reasonable to regard each NS less than 6 seconds as a non-replay at first. Given a NS with the frame sequence (nf_1, nf_2, …, nf_{Kn}), K_n is the total frame number. Let (nh_1, nh_2, …, nh_{Kn}) be the corresponding color histograms of K_n frames in the NS. Two histogram-based frame differences are defined by

$$DH_1(nh_i) = \frac{1}{8\times 8\times 8}\sum_{r=0}^{7}\sum_{g=0}^{7}\sum_{b=0}^{7}\left|nh_i(r,g,b) - nh_{i-1}(r,g,b)\right|, \quad (5)$$

$$DH_{15}(nh_i) = \frac{1}{8\times 8\times 8}\sum_{r=0}^{7}\sum_{g=0}^{7}\sum_{b=0}^{7}\left|nh_i(r,g,b) - nh_{i-15}(r,g,b)\right|. \quad (6)$$

Note that here we quantize each of r, g, and b into 8 levels, and DH_1 is used to measure the histogram difference of two neighboring frames, DH_{15} is for two frames with distance 15. After calculating $DH_1(nh_i)$ and $DH_{15}(nh_i)$ for each frame in NS, let σ_{DH1}(NS) and σ_{DH15}(NS) represent the standard deviations of sequences $DH_1(nh_i)$ and $DH_{15}(nh_i)$, respectively. Then these two standard deviations are considered as global variation features of a NS. For each NS, the two global variation features are used in k-means clustering to classify replay or non-replay. The technique detail can be referred to [19].

5. EXPERIMENTAL RESULTS

Our experiments are conducted by 4 NBA basketball videos, and all of them are broadcasted with semi-transparent scoreboards. All 4 semi-transparent scoreboard templates are extracted successfully. As to semantic event extraction, the webcast text is acquired from ESPN website. After annotating all basketball videos, the result is
evaluated by watching them with human eyes. An event is detected as a hit if the manually generated event boundary is covered by the proposed method. As can be seen in Table 1, the detection rate of annotation result without video frames partition is horrible. On the contrary, the detection rate of the proposed method reaches 100%. The main reason is that video frames partition can prevent unnecessary game clock recognition from frames without a scoreboard and raise the digit recognition rate. This also solves the challenge of discontinuity in basketball games.

As to slow motion replay detection, in order to explain the performance of k-means clustering, the two global features of each non-scoreboard segment in one of the experimented basketball videos are shown in Fig. 5. From this figure, we can see that replay segments and non-replay segments can be well-separated based on the proposed two global features.

Table 2 shows the replay detection results. It can be seen that the proposed method reaches extremely high recall rate to fulfill the expectation of highlight generation and video summarization. The rare false alarms are acceptable because they are all game-related video segments.

Table 1. Semantic events extraction results of the proposed framework.

Match	Annotation without video frames partition	Annotation with video frames partition
	Correctly detected number / Total event number (Detection rate)	
BOS-ORL (09/05/09)	57/169 (33.7%)	169/169 (100%)
LAL-DEN (23/05/09)	132/179 (73.7%)	179/179 (100%)
DEN-LAL (28/05/09)	136/182 (74.7%)	182/182 (100%)
LAL-DEN (30/05/09)	69/213 (32.4%)	213/213 (100%)
Average	**53.0%**	**100%**

Table 2. Replay detection results of the proposed framework.

Match	Correctly Detected	Total Detected	Precision	Total Replays	Recall
BOS-ORL (09/05/09)	22	31	71%	22	100%
LAL-DEN (23/05/09)	48	49	98%	48	100%
DEN-LAL (28/05/09)	44	45	98%	44	100%
LAL-DEN (30/05/09)	33	36	92%	33	100%
Total	**147**	**161**	**91%**	**147**	**100%**

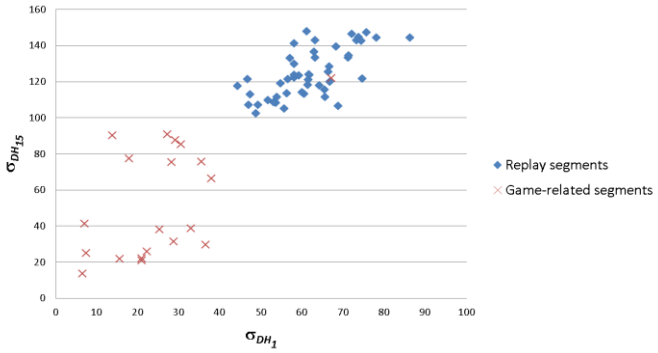


Fig. 5. The two global features of each NS in a basketball video.

6. CONCLUSIONS

In this paper, a novel framework for sports video analysis, which provides flexibility to combine different schemes of event extraction and those of replay detection, is proposed. Two semantic resource extraction schemes are introduced and incorporated in our framework to tackle challenges of basketball video analysis. The novelty of video frames partition prevents semantic resource extraction from a lot of unnecessary processing frames, so the performance and detection rate can be increased. Since no basketball specific feature is used in our work, it is expectable to push forward the framework to other sports videos in near future.

7. ACKNOWLEDGMENT

This work is supported in part by National Science Council of Republic of China under grant NSC-100-2221-E-009-140-MY2.

8. REFERENCES

[1] Y. H. Chen and L. Y. Deng, “Event mining and indexing in basketball video,” in *Int. Conf. on Genetic and Evolutionary Computing (ICGEC)*, pp. 247-251, Aug. 29-Sep. 1, 2011.

[2] E. Hassan, S. Chaudhury, M. Gopal, and V. Garg, “A hybrid framework for event detection using multi-modal features,” in *Int. Conf. on Computer Vision Workshops (ICCV Workshops)*, pp. 1510-1515, 2011.

[3] H. G. Kim and J. H. Lee, “Indexing of player events using multimodal cues in golf videos,” in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, pp. 1-6, 2011.

[4] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, “Personalized abstraction of broadcasted American football video by highlight selection,” *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 575–586, 2004.

[5] H. Xu and T. Chua, “Fusion of audio-visual features and external knowledge for event detection in team sports video,” in *Proc.Workshop on Multimedia Information Retrieval (MIR ’04)*, pp. 127-134, 2004.

[6] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, “Live sports event detection based on broadcast video and web-casting text,” in *Proc. ACM Int. Conf. on Multimedia (MM ‘06)*, pp. 221-230, 2006.

[7] C. Xu, J. Wang, H. Lu, and Y. Zhang, “A novel framework for semantic annotation and personalized retrieval of sports video,” *IEEE Trans. Multimedia*, vol. 10, no. 3, pp.421–436, 2008.

[8] C. Xu, Y. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, “Using webcast text for semantic event detection in broadcast sports video,” *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1342-1355, 2008.

[9] P. Lin, S. Li, T. Tsai, and Y. Tsai, “Tagging webcast text in baseball videos by video segmentation and text alignment,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 999-1013, 2012.

[10] H. Pan, B. Li, and M. I. Sezan, “Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions,” in *Proc. ICASSP’02*, pp. 3385-3388, 2002.

[11] L. Y. Duan, M. Xu, Q. Tian, and C. S. Xu, “Mean shift based video segment representation and applications to replay detection,” in *Proc. ICASSP’04*, pp. 709-712, 2004.

[12] Q. Huang, J. M. Hu, W. Hu, T. Wang, H. L. Bai, and Y. M. Zhang, “A reliable logo and replay detector for sports video,” in *Proc. ICME’07*, pp. 1695-1698, 2007.

[13] N. Nguyen and A. Yoshitaka, “Shot type and replay detection for soccer video parsing," in *Proc. ISM’12*, pp. 344-347, 2012.

[14] X. L. Zhang and M. Zhi, “Slow motion replay detection of tennis video based on color auto-correlogram,” in *Proc. ICDIP’12*, pp. 83341C, 2012.

[15] F. Zhao, Y. Dong, Z. Wei, and H. L. Wang, “Matching logos for slow motion replay detection in broadcast sports video,” in *Proc. ICASSP’12*, pp. 1409-1412, 2012.

[16] E. J. Farn, L. H. Chen, and J. H. Liou, “A new slow-motion replay extractor for soccer game videos,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol.17, no. 8, pp.1467-1481, 2003.

[17] L. Wang, X. Liu, S. Lin, G. Xu, and H. Y. Shum, “Generic slow-motion replay detection in sports video,” in *Proc. ICIP’04*, pp. 1585-1588, 2004.

[18] B. Han, Y. Yan, Z. H. Chen, C. Liu, and W. G. Wu, “A general framework for automatic on-line replay detection in sports video,” in *Proc. MM’09*, pp. 501-504, 2009.

[19] C. M. Chen and L. H. Chen, "A novel method for slow motion replay detection in broadcast basketball video," *Multimedia Tools and Applicat.*, submitted.

[20] C. M. Chen and L. H. Chen, “A novel approach for semantic event extraction from sports webcast text,” *Multimedia Tools and Applicat.*, to be published.