COVER SHEET

Multi-level Semantic Analysis for **Sports Video**

Dian W. Tjondronegoro [1], Yi-Ping Phoebe Chen [2]

[1] School of Information Systems, Queensland University of Technology
dian@qut.edu.au
[2] School of Information Technology, Deakin University,
Melbourne, Australia
phoebe@deakin.edu.au

Abstract. There has been a huge increase in the utilization of video as one of the most preferred type of media due to its content richness for many significant applications including sports. To sustain an ongoing rapid growth of sports video, there is an emerging demand for a sophisticated content-based indexing system. Users recall video contents in a high-level abstraction while video is generally stored as an arbitrary sequence of audio-visual tracks. To bridge this gap, this paper will demonstrate the use of domain knowledge and characteristics to design the extraction of high-level concepts directly from audio-visual features. In particular, we propose a multi-level semantic analysis framework to optimize the sharing of domain characteristics.

1 Introduction

Sports video indexing approaches can be categorised based on the two main levels of video content: low-level (perceptual) *features* and high-level *semantic* annotation [1-4]. The main benefits of *feature-based indexing* techniques: 1) they can be fully automated using feature extraction techniques such as image and sound analysis and 2) users can do similarity search using certain feature characteristics such as the shape and colour of the objects on a frame or the volume of the sound track. However, feature-based indexing tends to ignore the semantic contents whereas users mostly want to search video based on the semantic rather than the low-level characteristics. Moreover, there are some elements beyond perceptual level (widely known as the *semantic gap*) which can make low-level feature based indexing tedious and inaccurate. For example, users cannot always describe the visual characteristics of certain objects they want to view for each query. In contrast, the main benefit of *semantic-based indexing* is the ability to support more natural queries. However, semantic annotation is still generally time-consuming, and often incomplete due to limitations of manual supervision and the currently available techniques for automatic semantic extraction.

In order to bridge the semantic gap between low-level features and high-level semantic, a mid-level feature layer can be introduced [5]. The main purpose is to separate (sports) specific knowledge and rules from the low-level feature extraction; thus making it less domain-specific and robust for different sports. However, the use of mid-level features has not been optimized since most of the current semantic detection schemes still rely heavily on domain knowledge (see [6] as example).

It should be noted that some sport events can be detected using a more generalized knowledge while some have to use very specific rules. Thus, rather than immediately extracting domain-specific events, we can firstly detect generic key events based on, for example, replay scene, excited crowd/commentator, and whistle. Thus, stretching the semantic layer will allow us to generalize common properties of sports video, while minimizing the use of domain knowledge as far as possible. The longer that we can stay away from committing into a particular sport domain, the more robust that the designed algorithms and framework would be.

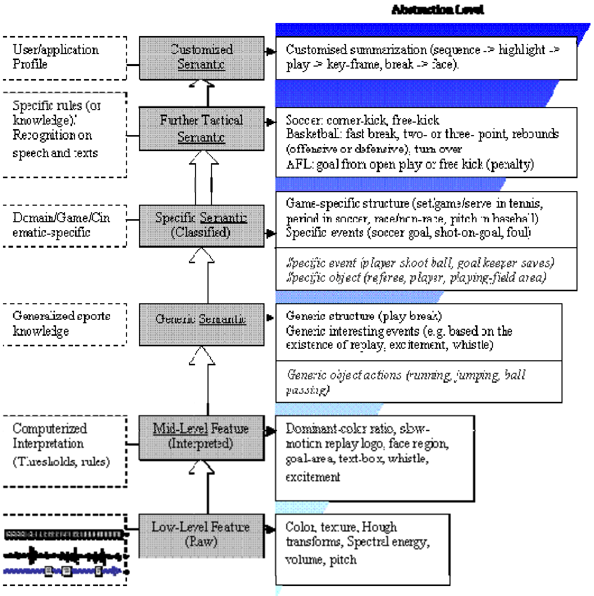
Abstraction Level
User/application Profile
Customized Semantic
Customised summarization (sequence -> highlight -> play -> key-frame, break -> face).
Specific rules (or knowledge)/ Recognition on speech and texts
Further Tactical Semantic
Soccer: corner-kick, free-kick
Basketball: fast break, two- or three- point, rebounds (offensive or defensive), turn over
AFL: goal from open play or free kick (penalty)
Domain/Game/Cinematic-specific
Specific Semantic (Classified)
Game-specific structure (set/game/serve in tennis, period in soccer, race/non-race, pitch in baseball)
Specific events (soccer goal, shot-on-goal, foul)
Specific event (player shoot ball, goal keeper saves)
Specific object (referee, player, playing-field area)
Generalized sports knowledge
Generic Semantic
Generic structure (play break)
Generic interesting events (e.g. based on the existence of replay, excitement, whistle)
Generic object actions (running, jumping, ball passing)
Computerized Interpretation (Thresholds, rules)
Mid-Level Feature (Interpreted)
Dominant-color ratio, slow-motion replay logo, face region, goal-area, text-box, whistle, excitement
Low-Level Feature (Raw)
Color, texture, Hough transforms, Spectral energy, volume, pitch

Fig. 1. Multi-level Semantic Analysis in Sports Video Framework

Hence, we propose to extend semantic layer (illustrated in Figure 1) by dividing *semantic features* into:

– *Generic semantic*. This layer can be detected using generalized knowledge on the common features shared by most of sports videos. For example, *interesting events* in sport videos are generally detectable using generic mid-level features such as whistle, excited crowd/commentator, replay scene, and text display.
– *Specific/classified semantic*. This layer is usually detected using a set of more specific characteristics in different sports based on domain-knowledge. The main purpose is to classify interesting events into domain-specific events such as soccer goal in order to achieve a more sophisticated retrievals and browsing.
– *Further tactical semantic*. This layer can be detected by focusing on further-specific features and characteristics of a particular sport. Thus, tactical semantic need to be separated from specific semantic due to its requirement to use more complex and less-generic audio-visual features. For example, *corner kick* and *free-kick* in soccer needs to be detected using specific analysis rules on soccer-field and

player-ball motion interpretation (such as in [7]). Playing-field analysis and motion interpretation are less generic than excitement and break shot detection. For instance, excitement detection algorithm can be shared for different sports by adjusting the value of some parameters (or thresholds), whereas the algorithm to detect corner (playing-field) in soccer will not be applicable for tennis due to the difference between soccer and tennis fields.

- *Customized semantic*. This layer can be formed by selecting particular semantics using user or application usage-log and preferences. For example, a sport summary can be structured using an integrated highlights and play-breaks using some text-alternative annotations [8].

The following describes an example of how the semantic layers are connected. During the extraction of generic events, play-breaks which can contain interesting events are detected. Statistics of mid-level features in each play-break can be used to construct heuristic-rules that classify the events contained into specific domain such as soccer *goal* (i.e. each play-break contain zero/more highlight events). When soccer goal is detected; if the play scene switch from left-goal-area to right-goal-area in less than 3-4 seconds, then further-tactical event of *counter-attack* is detected. Customized semantic is constructed by combining particular events based on user/application requirements.

One of the main benefits from using the proposed extended semantic layer is to achieve clearer boundaries between semantic analysis approaches; thus enabling a more systematic framework for designing event detection. For example, we can categorize goal detection in soccer [6] and basketball [9] into ‘specific-semantic analysis’, whereas the work presented to detect corner kick and free-kick in soccer [7] should be categorized under ‘further-tactical events analysis’. Using this framework, we can achieve more organized reviews, benchmark, and extensions from current solutions. This paper focuses on the extraction of generic and specific semantic.

2. Generic Semantic Extraction

Play-Break

A sport video is usually started with a pre-game scene which often contains introductions, commentaries and predictions. After a game is started, the game is played until there is an event which causes break. After a play being stopped, it will be resumed until being stopped again. This play-break phase is iterative until the end of the game which is then followed by a post-game scene which has similar content to pre-game.

Duration of camera-views have been used successfully for play-break segmentation (such as in [6]) since a long global shot (interleaved shortly by other shots) is usually used by broadcasters to describe attacking play which could end because of a goal. After a goal is scored, zoom-in and close-up shots will be dominantly used to capture players and supporters celebration. We have extended this approach by adding replay-based correction to design the algorithm for play-break segmentation which is effective and reliable for soccer, AFL and basketball video.

Figure 2 shows the various scenarios on how a replay scene can fix the boundaries of play-break sequences. In addition to these scenarios, it is important to note that most broadcasters would play a long replay scene(s) during a long break, such as goal celebration in soccer or timeout in basketball, in order to keep viewers' attention. However, some broadcasters insert some advertisements (*ads*) in-between or during the replay. Thus, to obtain the correct length of the total break, these types of ads should not be removed; at least the total length of the ads has to be taken into account.

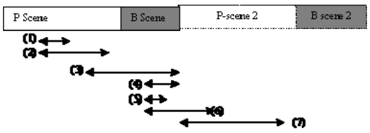

Fig. 2. Locations of Replays in Play-breaks

Based on these scenarios, an algorithm to perform replay-scene based play-break segmentation has been developed.

```
(1) If [R strict_during P] & [(R.e – P.e) >= dur_thres]
      B.s = R.s; B.e = R.e; Create a new sequence where [$P_2$.s = R.e+1] & [$P_2$.e P.e]
(2) If [R strict_during P] & [(R.e – P.e) <= dur_thres]
      P.e = R.e; B.s = R.e+1
(3) If [R meets B] & [R.s < P.e]
      P.e = R.s
(4-5) If [R during B] & [R meets B] ) OR (If [R strict_during B])
      No processing required
(6) If  [R during B] & [(R.e – $P_2$.s) >= dur_thres]
      B.e = R.e; Amend the neighbor sequence: [$P_2$.s = R.e+1]
(7) If [R during $P_2$ ]  & [(R.e – $P_2$.s) >= dur_thres]
      Attach sequence 2 to sequence 1 (i.e. combine seq 1 and seq 2 into 1 sequence)
```

where, If A *strict_during* B, (A.s > B.s) & (A.e < B.e); if A *during* B, (A.s > B.s & A.e <= B.e) OR (A.s >= B.s & A.e < B.e); If A *meets* B, A.e = B.e.

Table 1 depicts the performance of the play-break segmentation algorithm on 2 soccer videos which were recorded from Champions League: Milan-Depor (S1)**,** Madrid-Bayern (S2). S1-1 is video1 first half while S1-2 is video2 second half. In this table, *RC*=Replay-based Correction for play-break segmentation, *PD*=perfectly detected, *D*=detected, *M*=missed, *F*=false detection, *Tr* =Total number in Truth, *Det* =Total Detected, *RR*=Recall Rate, *PR*=Precision Rate, and *PDdecr*=perfectly detected decrease rate if *RC* is not used; where: *Tru=PD+D+M*, *Det=PD+D+F*, *RR=*(*PD+D+M*)/*Tru**100%, *PR=*(*PD+D*)/*Det**100%, and *PD_Decr=*(PD-D)/PD*100%.

The results have confirmed that *RC* improves the play-break segmentation performance. It is due to the fact that many (if not most) replay scenes use global (i.e. play) shots. This is shown by all *PD_decr, RR,* and *PR* since *RC* always improves all of these performance statistics. In particular, the *RR* and *PR* for soccer 1-1 with *RC* are 100% each but they are reduced to below 50% without *RC*. In soccer 1-1 without *RC*, the *PD* dropped from 49 to 12 (i.e. 75% worse) whereas *M* increases from 0 to 25 and *F* increases from 0 to 5. This is due to the fact that soccer1 video contains many replay scenes which are played abruptly during a play, thereby causing a too-long play scene and missing a break.

Generic Key Events

In Table 2, we have demonstrated the benefits of using three features: whistle, excitement and text, in order to detect generic key events (see [10] for more details). Based on this table, it is clear that whistle detection is very fast, but it can only localize 20 to 30% of the total highlights which are mostly caused by foul and offside. By combining whistle and excitement, users only need to wait slightly longer to detect 80% to 90% of the highlights since excitement can locate most of exciting events, such as *good display of attacking/defending, goal,* and *free kick.* Excitement detection is particularly effective to localize goal highlights due to the massive amount of excitement during the start of a good attack which often leads to the goal itself. When whistle and text detection are combined, the number of highlights detected will only slightly increase while the waiting-period is longer than using excitement. This is due to the fact that visual features are generally more expensive computationally than audio features. Text detection is needed to localize *start of a match*, *goal* and *shot on goal*, as well as confirming offside and foul events.

3. Classified Semantic

While generic key events are good for casual video skimming, domain-specific (or classified) highlights will support more useful browsing and query applications. Unlike the current heuristic rules scheme for highlight detection, our system does not use manual knowledge. This is achieved by using a semi-supervised training on 20 samples from different broadcasters and matches for each highlight to determine the characteristics of play-breaks containing different highlights and no highlights. It is semi-supervised as we manually classify the specific highlight that each play-break sequence (for training) contains. Moreover, the automatically detected play-break boundaries and mid-level features locations within each play-break (such as excitement) are manually checked to ensure the accuracy of training.

We have demonstrated that the statistics of mid-level features (e.g. close-up and excitement ratio) within a play-break sequence can be used for classification of sport domain events. At this stage we have successfully detected highlights and classify them into: goal, shoot, foul, or non in soccer; goal, behind, mark, tackle, or non in AFL (see [11] for soccer and AFL experiment). In this section, we will discuss the experiment to detect *goal*, *free-throw, foul,* or *timeout* in basketball.

Using the trained statistics, heuristic rules for basketball events can be identified:

- *Timeout*: compared to any other highlight, this event has the longest possible duration (of play-break), mostly containing break shots. Consequently, there are many slow motion replays and close-up shots to keep viewers’ interest. This event usually contains the least portion of near-goal play.
- *Goal/Free throw*: compared to foul, goal and free throw contain a large amount of near goal and replay scene is never played. To differentiate goal from free throw, the statistics show that goal contains less close-up shots; thus play shot is more dominant. However, free throw usually contains longer near-goal and duration.

– *Foul:* when a sequence is less likely to contain goal, free throw or goal, foul can usually be detected. Moreover, the close-up is more than goal and free throw but less than timeout; play and break ratio is almost equal; at least one replay scene; and the excitement is less than goal but more than free throw.

The statistical-driven heuristic rules have been tested to show its effectiveness and robustness using a large dataset of basketball videos (as described in Table 3) where each sport is recorded from different broadcasters, competition, match, and/or stage of competition. Table 4 depicts the performance of highlight events classification in basketball videos. The experimental results show high *RR* and reasonable *PR*.

Table 1. Play-break Segmentation Performance in Soccer Videos

	Soccer Play-break detection								
Video	*PD*	*D*	*M*	*F*	*Tru*	*Det*	*RR*	*PR*	*PD_decr*
S1-1 (*RC*)	49	0	0	0	49	49	100.00	100.00	
S1-1	12	12	25	5	49	54	48.98	44.44	75.51
S1-2(*RC*)	53	0	0	1	53	54	100.00	98.15	
S1-2	36	10	7	1	53	54	86.79	85.19	32.08
S2-1(*RC*)	54	1	1	12	56	68	98.21	80.88	
S2-1	53	2	1	12	56	68	98.21	80.88	1.85
S2-2(*RC*)	58	1	0	7	59	66	100.00	89.39	
S2-2	55	4	0	7	59	66	100.00	89.39	5.17

Table 2. Generic Events Detection Results (#H: number of highlights, Time is in minute)

Sample Video	**Total High-lights**	**Automatically Detected Highlights**							
		Whistle Only (W)		**Whistle + Text (W+T)**		**Whistle + Excitement (W+E)**		**W+T+E**	
		#H	Time	#H	Time	#H	Time	#H	Time
Soccer1 (40mins)	**62**	**11**	1.7	**13**	37.1	**54**	22.9	**56**	58.2
Soccer2 (20mins)	**24**	**7**	0.7	**8**	24.8	**22**	10.6	**23**	35.4
Soccer3 (20mins)	**40**	**11**	0.7	**11**	26.7	**39**	8.8	**39**	35.5
Soccer4 (20 mins)	**22**	**2**	0.9	**3**	18.1	**21**	8.9	**22**	19
Rugby1 (20 mins)	**34**	**18**	0.9	**20**	20.6	**25**	10.9	**27**	29.9
Rugby2 (17 mins)	**21**	**8**	0.7	**9**	16.9	**18**	9.6	**19**	18.5
Basketball (15 m)	**37**	**7**	0.8	**12**	14.6	**30**	7.9	**35**	21.9
Tennis (20 mins)	**40**	**0**	0	**0**	0	**33**	9.9	**33**	28.8
Netball (19 mins)	**43**	**36**	0.4	**39**	8.8	**38**	4.9	**41**	13.4
Average time spent for 1 min segment		**0.04**		**1.06**		**0.52**		**1.49**	

Table 3. Basketball Video Samples

Group (broadcaster)	**Basketball Videos "team1-teams2_period-[duration]"**
Athens 2004 Olympics (Seven)	Women: AusBrazil_ 1,2,3-[19:50,19:41,4:20] Women: Russia-USA_3-[19:58] Men: Australia-USA_1,2-[29:51,6:15]
Athens 2004 Olympics (SBS)	Men: USA-Angola_2,3-[22:25,15:01] Women: Australia-USA_1,2-[24:04-11:11]

Table 4. Highlight Classification in Basketball Videos

Ground truth	Highlight classification of 5 basketball videos				
	Goal	Free throw	Foul	Timeout	Truth
Goal	**56**	0	0	2	58
Free throw	4	**14**	0	0	18
Foul	21	2	**28**	3	54
Timeout	0	0	0	**13**	13
Total Detected	81	16	28	18	

4. Conclusion and Future Work

In this paper, we have presented a multi-level semantic analysis framework for sports video by extending semantic layer into: generic-, classified-, further-tactical- and customized-semantic. This paper focuses on the analysis methods and performance results for extraction of generic and classified semantic. In future work, we aim to enhance the performance of the semantic analysis while extending the scope of detectable mid-level features and semantic for various sports genre. We also aim to demonstrate the extraction of further-tactical semantic such as soccer free kick.

References

1. Lu, G.J.: Multimedia database management systems. Artech House. Boston London (1999).
2. Tusch, R., Kosch H., Böszörményi, L.: VIDEX: an integrated generic video indexing approach. Proceedings of the 8th ACM international conference on Multimedia. ACM Press. Marina del Rey California United States (2000) 448-451.
3. Elgmagarmid, A.K. (ed): Video database systems: issues, products, and applications. Kluwer Academic Publishers. Boston London (1997).
4. Djeraba, C.: Content-based multimedia indexing and retrieval. IEEE Multimedia. Vol. 9(2) (2002) 18-22.
5. Duan, L.-Y., Min, X., Chua, T-S., Qi, T., Xu, C-S: A mid-level representation framework for semantic sports video analysis. Proceedings of the 11th ACM international conference on Multimedia. ACM Press. Berkeley USA (2003) 33-44.
6. Ekin, A., Tekalp, M.: Automatic Soccer Video Analysis and Summarization. IEEE Transaction on Image Processing. Vol. 12(7) (2003) 796-807.
7. Han, M., Hua, W., Chen, T., Gong, Y.: Feature design in soccer video indexing. Proceedings of the 4th Conference on Information, Communications and Signal Processing. IEEE. Singapore (2003) 950-954.
8. Tjondronegoro, D., Chen, Y-P.P., Pham, B.: The Power of Play-Break for Automatic Detection and Browsing of Self-consumable Sport Video Highlights. The Proceedings of the 6th International Multimedia Information Retrieval Workshop. ACM Press. New York USA (2004) 267-274.
9. Nepal, S., Srinivasan, U., Reynolds, G.: Automatic detection of goal segments in basketball videos. Proceedings of the 9th ACM International Conference on Multimedia. ACM Press. Ottawa Canada (2001) 261-269.
10. Tjondronegoro, D., Chen, Y-P.P., Pham, B.: Integrating Highlights to Play-break Sequences for More Complete Sport Video Summarization. IEEE Multimedia. Vol.11(4) (2004) 22-37.
11. Tjondronegoro, D., Chen, Y-P.P., Pham, B.: A Statistical-driven Approach for Automatic Classification of Events in AFL Video Highlights. Proceedings of the 28th Australasian Computer Science Conference. ACS. Newcastle Australia (2005) 209-218.