USING DECISION-TREE TO AUTOMATICALLY CONSTRUCT LEARNED-HEURISTICS FOR EVENTS CLASSIFICATION IN SPORTS VIDEO

Dian Tjondronegoro[1] *Yi-Ping Phoebe Chen*[2]

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

ABSTRACT

Automatic events classification is an essential requirement for constructing an effective sports video summary. It has become a well-known theory that the high-level semantics in sport video can be "computationally interpreted" based on the occurrences of specific audio and visual features which can be extracted automatically. State-of-the-art solutions for features-based event classification have only relied on either manual-knowledge based heuristics or machine learning. To bridge the gaps, we have successfully combined the two approaches by using learning-based heuristics. The heuristics are constructed automatically using decision tree while manual supervision is only required to check the features and highlight contained in each training segment. Thus, fully automated construction of classification system for sports video events has been achieved. A comprehensive experiment on 10 hours video dataset, with five full-match soccer and five full-match basketball videos, has demonstrated the effectiveness/robustness of our algorithms.

1. INTRODUCTION

The main components of a video document are *semantic content* and *audiovisual* (*AV*) *presentation*. S*emantic content* is the idea, knowledge, story, message or entertainment conveyed by the video data. It is the most complex part of video data as the semantic information of video can be expressed either implicitly or explicitly. Viewers need to apply their knowledge to understand the implicit semantic after seeing or hearing the *audiovisual* presentation whereas they should be able to understand semantic more intuitively. An example of explicit semantic is the text displays in sports video to inform viewers of the current score board. Similarly to the natural process of acquiring implicit semantic information, sports events can be automatically detected based on the occurrences of specific audio and visual features. To date, there are two main approaches to fuse audio-visual features for semantic extraction. One alternative is to use manual heuristic rules. For example, the temporal gaps between specific features during basketball goal have a predictable pattern that can be perceived manually [1]. The main benefit of this approach is the absence of comprehensive training for each highlight and the computations are relatively less complex. However, this method usually relies on manual observations to construct the detection models for different events. Even though the numbers of domains and events of interest are limited and the amount of efforts is affordable, we should aim to reduce the subjectivity and limitation of manual decisions.

Another alternative, called *machine-learning* approach, uses probabilistic models to automatically capture the unique patterns of audio visual feature-measurements in specific (highlight) events. For example, Hidden Markov Model (HMM) can be trained to capture the transitions of 'still, standing, walking, throwing, jumping-down and running-down' states during athletic sports' events, which are detected based on color, texture and global-motion measurements [2]. The main benefit of using such approach is the potential robustness, thanks to the modest usage of domain-specific knowledge which is only needed to select the best features set to describe each event. However, one of the most challenging requirements for constructing reliable models is to use features that can be detected flawlessly during training due to the absence of manual supervision. Moreover, HMM tries to capture the pattern of observations in a continuous time-period. This makes HMM too sensitive to noises and errors in features extraction.

Both of the above-mentioned alternatives still have two major drawbacks, namely, 1) the lack of a definitive

1-4244-0367-7/06/$20.00 ©2006 IEEE

solution for the scope of highlight detection such as where to start and finish the extraction. For example, Ekin et al [3] detect goals by examining the video-frames between the global shot that causes the goal and the global shot that shows the restart of the game. However, this template scope was not used to detect other events. On the other hand, Han et al [4] used a static temporal-segment of 30-40 sec (empirical) for soccer highlights detection. 2) The lack of a universal set of features for detecting different highlights and across different sports. Features that best describe a highlight are selected using domain knowledge. For instance, whistle in soccer is only used to detect foul and offside, while excitement and goal-area are used to identify goal attempt [5].

In this paper, we will present a novel attempt to bridge the two approaches by using learned-based heuristics. Our approach utilizes standard scope of detection and set of features for different events and sports domain. The heuristics are constructed automatically using decision tree. During training, minimum manual supervision is required to check the features and highlight contained in each training segment. Thus, rapid and fully automated construction of classification system for sports video events has been achieved. A comprehensive experiment on 10 hours dataset of sports videos, including soccer and basketball, has demonstrated the effectiveness and robustness of our algorithms.

2. FRAMEWORK OF EVENTS DETECTION

A *play* scene in sports video is when the game is flowing which can be stopped (i.e. become a *break* scene) due to various reasons such as goal and foul. Most broadcasted sport videos use transitions of typical shot types to emphasize story boundaries while aiding important contents with additional items. For example, a long global shot is normally used to describe an attacking *play* that could end with scoring of a goal. After a goal is scored, zoom-in and close-up shots will be dominantly used to capture players and supporters celebration during the *break*. Subsequently, some slow-motion replay shots and artificial texts are usually inserted to add some additional contents to the goal highlight. Based on this example, it should be clear that play-break sequences should be effective containers for a semantic content since they contain all the required details. Using this assumption, we should be able to extract all the phenomenal features from play-break that can be utilized for highlights detection. Thus, as shown in Figure 1, the scoping of highlight (event) detection should be from the last play-shot until the last break shot.

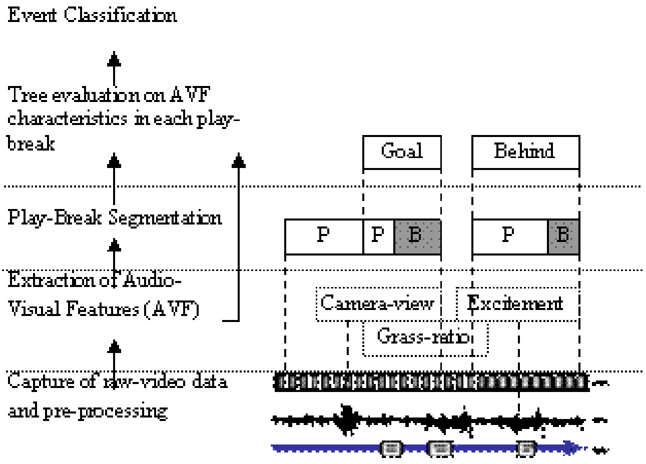
Event Classification
Tree evaluation on AVF characteristics in each play-break
Goal
Behind
Play-Break Segmentation
P
P
B
P
B
Extraction of Audio-Visual Features (AVF)
Camera-view
Excitement
Grass-ratio
Capture of raw-video data and pre-processing

Figure 1. Extracting Events from Play-Break.

3. AUTOMATIC CONSTRUCTION OF LEARNING-BASED HEURISTICS FOR EVENTS DETECTION

Learning is performed by performing some manual interventions. First is correcting the boundaries of each detected play-break (PB) segment. Second is labeling the specific event contained in each PB. Third is correcting the audio visual features occurrence, such as boundaries of replay scene.

For each event (to be classified), learning is performed based on the following parameters:

- D = duration of currently-observed play-break sequence.
- B = duration of break / D.
- P = duration of play scene / D.
- R = duration of (slow-motion) replay scene in the sequence. This measurement implicitly represents the number of slow motion replay shots which is generally hard to be determined due to many camera changes during a slow motion replay.
- E = duration of excitement / D. Typically, highlight events consist of higher excitement ratio whereas non-highlight usually contain no excitement.
- N = duration of the frames containing goal-area / D. A high ratio of near goal area during a play potentially indicates goal or goal-attempt.
- C = length of close-up views that includes crowd, stadium, and advertisements within the sequence / D.

This set of features is selected as they are generally effective for describing sport events, in particular, soccer and basketball and any sports with similar characteristics. It should be noted that whistle occurrence was not used even though it is very useful for many sports; it is due to the fact that whistles are hardly audible and often falsely detected from whistle blown by audience. Similarly, inserted texts occurrence is not used as their location within a sequence is not predictable. For example, caption for a goal is usually displayed in the next play shot after goal celebration while caption for a shot is usually displayed during the break.

It should be noted that readers should refer to our earlier paper [6] to get algorithms and performance of our features extraction techniques, including play-break segmentation and detection of whistle, excitement, and goal-area. In this paper, we will only focus on the events classification.

In Table 1, we have provided an example of our training data for basketball foul event. The training data for all events in a particular sport domain is used to construct *tree-based classification model* that can predict the response (i.e. event) as a function of predictors (i.e. features). We have used the decision-tree tools provided by MATLAB 7 for our experiment. There are two some parameters that can be adjusted during training which can produce better/worse classification performance results:

- *prior*, prior probabilities of each event-class (e.g. in soccer, goal rarely happens compared to any other event)
- *split criterion* (*crit*), split criterion (i.e. what method of splitting)
- *split min* (*min*), minimum number of observations before a node is split into a tree-branch
- *prune*, whether pruning is performed on the tree

To get *prior*, we have used the actual number of events in the truth data from 5 full matches for each sports domain. In soccer: *prior* is calculated as (NH=502; Goal = 7/NH; Shot = 110/NH; Foul = 110/NH; Non = 275/NH), which means that out of 502 events, only 7 of them were goal. In basketball, (NH=143; Goal = 58/NH; FreeThrow = 18/NH; Foul = 54/NH; TimeOut = 13/NH). During experiment for each sports domain, we have tested different combinations of parameters that produce the best performance results (detailed discussion on the measurements will be discussed in the next section). The most-effective parameters were:

In soccer: crit = deviance, min = 20, prune = off

In basketball: crit = deviance, min = 15, prune = on.

Using 20 samples (from different matches and broadcasters) for each event, we have constructed reliable learned-heuristics for soccer and basketball, which are shown in Figure 2 and 3. Our system can classify *goal*, *foul*, shot-on-goal (*shot*), and non-interesting event (*non*) in soccer, as well as *goal*, *free-throw*, *foul*, *timeout*, and *non-interesting* event in basketball. These events were selected since they are often used to summarize soccer and basketball highlights. Moreover, non-interesting events are separately trained since they have distinctive characteristics too, just like the interesting events.

The main benefits of our learning approach are:

- To reduce noise in fully-unsupervised learning, construction of "correct" learning is optimized by minimum amount of manual correction
- Heuristics are constructed without the use of any domain knowledge
- Universal scope and set of measurement for event classification which is applicable for different sports

P	D (out of 2 mins)	excitement ratio	break ratio	R (out of 40 seconds)	N	C
0.35	0.17	0.70	0.65	0.15	0.29	0.35
0.58	0.10	0.00	0.42	0.00	0.00	0.50
0.53	0.28	0.26	0.47	0.18	0.22	0.24
0.37	0.34	0.24	0.63	0.30	0.13	0.22
0.54	0.11	0.54	0.46	0.00	0.00	0.00
0.64	0.21	0.56	0.36	0.00	0.38	0.24
0.64	0.30	0.78	0.36	0.00	0.26	0.31
0.26	0.23	0.30	0.74	0.23	0.29	0.15
0.28	0.30	0.19	0.72	0.20	0.70	0.22
0.58	0.22	0.15	0.42	0.18	0.93	0.08
0.38	0.24	0.14	0.62	0.00	0.64	0.14
0.45	0.18	0.23	0.55	0.00	0.50	0.14
0.62	0.22	0.58	0.38	0.00	0.75	0.62
0.64	0.12	0.21	0.36	0.00	0.89	0.43
0.13	0.50	0.27	0.87	0.00	0.25	0.57
0.63	0.33	0.33	0.38	0.20	0.28	0.00
0.33	0.25	0.43	0.67	0.18	0.30	0.27
0.72	0.38	0.26	0.28	0.33	0.67	0.11
0.65	0.22	0.65	0.35	0.23	0.47	0.12
0.31	0.13	0.00	0.69	0.28	0.60	0.69

Table 1. Sample of Training Data for Basketball Foul.

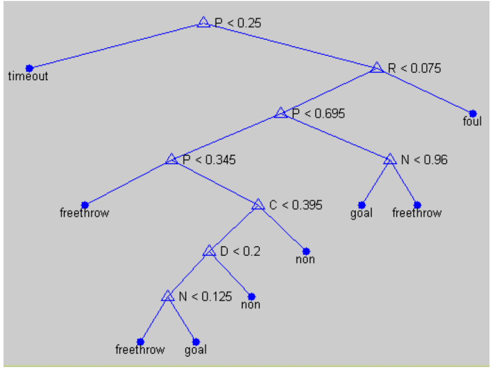

Figure 2. Decision Tree for Soccer Events Classification

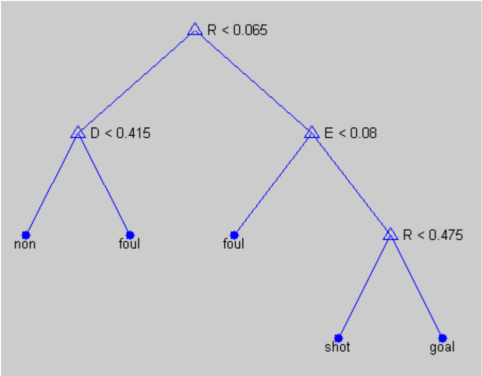

Figure 3. Decision Tree for Basketball Events Classification

5. EXPERIMENTAL RESULTS

Table 2 will describe the video samples used during experiment. For each sport, we have used videos from different competitions, broadcasters and/or stage of tournament. The purpose is, for example, final match is expected to contain more excitement than a group match while exhibition will show many replay scenes to display players' skills. For events classification, we manually developed the ground truth for each play-break sequence with the event contained. In order to measure the performance of highlights classification, we found that Recall (*RR*) and Precision Rate (*PR*) are not sufficiently accurate and expressive. The main reason is that we need to see precisely where the miss- and false-detections are. Therefore, we have provided the *RR, PR* and the actual detections results.

In Table 3 and 4, we have shown the details of detection results from 5 soccer videos, and 5 basketball video respectively. For example in Table 3, we have learned that from the 7 goal events, all 7 of them were detected correctly as goals. Whereas, out of 107 shots, 49 of them were correctly detected as shots, while 14 of them were detected as goals. This information is important, rather than mere recall and precision. For example, we can tolerate that shots and goals can be falsely classified between each other as they have closer relationship.

In Table 5, we have provided the Recall (*RR*) and Precision (*PR*) rate of each video samples. To calculate RR and PR, let *gg* to denote "truth goal, detected as goal", gs "truth goal, detected as shot", and so on. In soccer, the detection is calculated as follows:

$$Nc = gg + ss + ff + nn$$

$$Nf = (gs + gf) + (sg + sf) + (fg + fs) + (ng + ns + nf)$$

$$Nm = gn + sn + fn$$

Where, g = goal, s = shot, f = foul, n = non

In basketball,

$$Nc = gg+ss+ff+tt+nn$$

$$Nf = (gs + gf + gt) + (sg + sf + st) + (fg + fs + ft) + (tg + ts + tf) + (ng + ns +nf + nt)$$

$$Nm = gn + sn + fn + tn + fn$$

Where, g = goal, s = free-throw (shot on goal), f = foul, t = timeout, n = non

Thus,

$$RR = Nc/(Nc+Nm) * 100$$

$$PR = Nc/(Nc+Nf) * 100$$

During out experiment, we have emphasized on better Recall than Precision. In other words, we would prefer to get more interesting events, rather than to miss them (i.e. by classifying them as non-interesting). Likewise, a low PR can be tolerated since viewers will still get a generically interesting event, such as goal detected as foul.

Sample Group (Broadcaster)	Videos "team1-teams2_period-[duration]"
Soccer: UEFA Champions League Group Stage Matches (SBS)	*ManchesterUtd-Deportivo1,2-[9:51, 19:50]* *Madrid-Milan1,2[9:55,9:52]*
Soccer: UEFA Champions league (SBS) Elimination Rounds	*Juventus-Madrid1,2:[19:45,9:50]* *Milan-Internazionale1,2:[9:40,5:53]* Milan-Depor1,2-[51:15,49:36] **(S1)** Madrid-BayernMunich1,2-[59:41,59:00] **(S2)** Depor-Porto-[50:01,59:30] **(S3)**
Soccer: FIFA World cup Final (Nine)	*Brazil-Germany [9:29,19:46]*
Soccer: International Exhibition (SBS)	Aussie-SthAfrica1,2-[48:31,47:50] **(S4)**
Soccer: FIFA 100th Anniversary Exhibition (SBS)	Brazil-France1,2-[31:36,37:39] **(S5)**
Basketball: Athens 2004 Olympics (Seven)	Women: AusBrazil_ 1,2,3-[19:50,19:41,4:20] **(B1)** Women: Russia-USA_3-[19:58] **(B2)** Men: Australia-USA_1,2-[29:51,6:15] **(B3)**
Basketball: Athens 2004 Olympics (SBS)	Men: USA-Angola_2,3-[22:25,15:01] **(B4)** Women: Australia-USA_1,2-[24:04-11:11] **(B5)**

Table 2. Details of Sample Data for Experiments

Truth	Detected as				
	'goal'	'shot	'foul'	'non'	Total Truth
'goal'	7	0	0	0	7
'shot'	14	49	21	23	107
'foul'	16	41	31	24	112
'non'	2	15	57	167	241
Total Detected	39	105	109	214	

Table 3. Experimental Results on 5 Full-Match Soccer

Truth	Detected as					
	'goal'	'freethrow'	'non'	'foul'	'timeout'	Total Truth
'goal'	50	2	7	1	1	61
'freethrow'	4	8	3	0	1	16
'non'	10	4	11	6	0	31
'foul'	8	4	12	25	6	55
'timeout'	0	0	0	0	11	11
Total Detected	72	18	33	32	19	

Table 4. Experimental Results on 5 Full-Match Basketball

	RR	*PR*
S1	86.42	76.92
S2	80.00	80.00
S3	92.11	28.69
S4	84.72	62.89
S5	69.70	60.53
B1	92.00	57.50
B2	100.00	65.85
B3	92.86	65.00
B4	91.67	75.86
B5	86.96	55.56
Overall	87.64	62.88

Table 4. Recall and Precision of Experiments on each Video Sample

6. REFERENCES

[1] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of 'Goal' segments in basketball videos," presented at ACM International Conference on Multimedia, Ottawa; Canada, 2001.

[2] C. Wu, Y.-F. Ma, H.-J. Zhang, and Y.-Z. Zhong, "Events recognition by semantic inference for sports video," presented at Multimedia and Expo, 2002. Proceedings. 2002 IEEE International Conference on, 2002.

[3] A. Ekin and M. Tekalp, "Automatic Soccer Video Analysis and Summarization," *IEEE Transaction on Image Processing*, vol. 12, pp. 796-807, 2003.

[4] M. Han, W. Hua, T. Chen, and Y. Gong, "Feature design in soccer video indexing," presented at Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on, 2003.

[5] L.-Y. Duan, M. Xu, T.-S. Chua, T. Qi, and C.-S. Xu, "A mid-level representation framework for semantic sports video analysis," presented at ACM MM2004, Berkeley, USA, 2003.

[6] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham, "A Statistical-driven Approach for Automatic Classification of Events in AFL Video Highlights," presented at The 28th Australasian Computer Science Conference, Newcastle, Australia, 2005.