A Hierarchical Framework for Generic Sports Video Classification

Maheshkumar H. Kolekar and Somnath Sengupta

Electronics and Electrical Communication Engineering Department,
Indian Institute of Technology,
Kharagpur-721302, West Bengal, India
{mhkolekar, ssg}@ece.iitkgp.ernet.in

Abstract. A five layered, event driven hierarchical framework for generic sports video classification has been proposed in this paper. The top layer classifications are based on a few popular audio and video content analysis techniques like short-time energy and Zero Crossing Rate (ZCR) for audio and Hidden Markov Model (HMM) based techniques for video, using color and motion as features. The lower layer classifications are done by applying game specific rules to recognize major events of the game. The proposed framework has been successfully tested with cricket and football video sequences. The event-related classifications bring us a step closer to the ultimate goal of semantic classifications that would be ideally required for sports highlight generation.

1 Introduction

Event-based storage and retrieval of sports video sequences and automated generation of highlights are highly demanding topics, because of their popularity and commercial importance. Therefore, there has been a widespread studies in the field of sports video classification. E. Kijak et. al. [1] have presented the use of HMM for the structure analysis of Tennis video. J. Assfalg et. al. [2] have worked upon football video classification using camera motion and player's location. L. Duan et. al. [3] have proposed the color characterization model for sports video indexing and browsing. L. Xie et. al. [4] have proposed an algorithm for parsing the structure of soccer sports video. These works provide fairly compressive, solutions to the task outlined, however the challenge of developing a solution or scheme that can reveal common structures of multiple events across multiple domains remains under-investigated. In practice though, such a scheme could not exist without some limit of domain constraint, i. e. the design of common feature extraction metrics applied to two vastly different sports types. On the other hand it is important to avoid becoming too context specific. With this trade-off in mind, our research is aimed towards designing techniques such that they can be globally applied to all sports types, which come under the umbrella of 'ball and field sports'. Recently, unified general frameworks were proposed in [5],[6]. In their work, excitements were extracted for highlight generation, but

P.J. Narayanan et al. (Eds.): ACCV 2006, LNCS 3852, pp. 633–642, 2006.

detailed event classification was not carried out. The sports video classification schemes proposed till date fail to respond to action-based queries, such as "extract the goal clips out of this football sequence", or "find out when the batsman got run-out in this cricket video", etc. Such queries may be always needed for editing and retrieval.

Very recently, W. Hsu et. al. [7] have presented the scheme for the fusion of multimode features for TRECVID news video classification. L. Chaisorn et. al. [8] has proposed two-level framework for news classification. In these approaches, they segment the full video into video shots and then clustering the shots to generate the concept hierarchies. In contrast to these method, we have used the top down approach, which permits us to avoid clustering and consequently improves the classification accuracy and also maintains the temporal order of shots.

Successful solution of this problem has to address two basic issues; (a) the classification should be event-related, and (b) such events should fit well within a generic framework that can be used for any popular sports. The first issue is not easy to solve since the classification schemes do not use semantic information directly, but use basic clues like color, motion etc. At the top layers, we use these clues to achieve basic classifications and for the subsequent layers, we apply event-driven rules from the top layers to recognize the events. We successfully tried our approach with two popular games - football and cricket, but it is applicable to other games as well. To address the second issue, we have proposed a generic event representation framework, that is simple, hierarchical in nature and makes indexing and retrieval process easy, and straightforward.

The fundamental problem associated with the top-layer video classification is the large volume of data that we have to deal with. In every game, there are moments of excitement, with relatively dull periods in between. Only moments of excitements qualify for inclusion in the highlights and the dull periods, which are often lengthy in terms of number of frames, need to be filtered out. Excitements are always accompanied by significant audio content resulting from spectators' cheering, increase in the audio level of the commentators' voice etc. After carrying out large set of experiments, we conclude that audio serves as the most basic clue to filter out the dull contents and extract clips that may qualify for inclusion in highlights. We have used two popular audio content analysis techniques- short-time energy and ZCR for extracting possible moments of excitements. However, all excitements detected through audio features may not correspond to game excitements. Even commercial clips are sometimes associated with excitement type of audio contents and these must be detected and filtered out. In the next layers, video features, like color and motion have been used for classification. One of the major characteristics of sports video is that the sequences are highly structured in the sense that the number of events is usually limited in number and there are repetitive transitions, often back and forth between those events. Using a large number of video sequences for training, we have derived the scene structure in terms of the transition probabilities from one event to the other and trained a HMM model [9] for classification of the events.

2 Hierarchical Classifications

The tree diagram shown in Fig 1 is our proposed generic framework for the sports video classification. At the top layer, the requirement is to skim the video sequence to a significant extent and extract the possible moments of excitements. As explained in the previous section, audio features serve as a very important clue for this top-layer classification, which is essentially binary - excitement (L1: class-0) and non-excitement (L1: class-1). Of the clips labeled as "excitement" in level-1, some frames show direct game actions (L2: class-0), some display the spectators (L2: class-1), especially after the major events like goal in football and "out" or "sixer" in cricket. The first level audio classification even picks up the commercials (L2: class-2), since often those are presented with exciting tones and background effects added. At the next level, i.e., level-3, the game actions are further sub-classified into real-time events, post-event activities and replay. Every major event is followed by some post-event activities like players' celebrations and finally, replays are presented, while the audio excitements still continue. At level-4, real-time shots (L3: class-0) are classified into actions, based on a set of rules applied on real-time shots and post-event activities. At the next level, the actions are further classified into a set of rule-based sub-actions. The definition of action and sub-action can be specialized to a specific sports based on specific domain knowledge. For example, an action can be wicket and sub-action can be the type of wicket e.g. bowled, catch etc. in the cricket. In football, goal is an action and the type of goal, e.g. goal by penalty kick, goal by head etc. are sub-actions. The rules applied for action and sub-action detection are game-

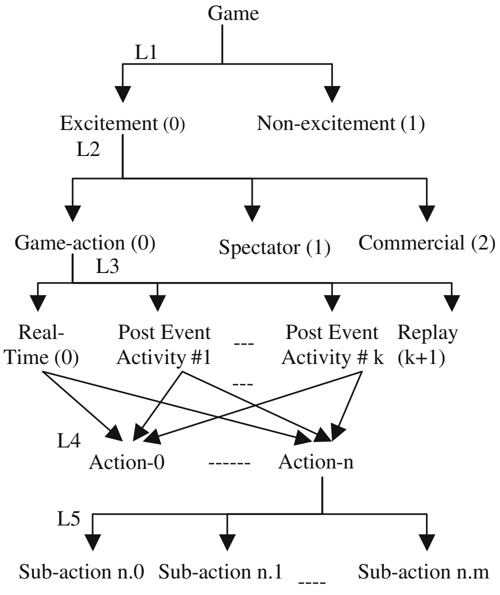

Fig. 1. Tree Diagram of Hierarchical Structure

specific. For example, in cricket, if the real-time actions are followed by fielders' celebrations and batsman's departure in close-up, it is a wicket, otherwise, it is a hit. In football, if the real-time actions are followed by players' celebrations and close-up, it is a goal, otherwise, a goal-miss.

3 Event Detection and Classification

3.1 Excitement Detection at Level 1

We have observed that whenever there is an important activity in the game, there is a corresponding increase in audio energy. We have used two popular audio content analysis techniques- short-time energy and ZCR [10] for extracting commercials. A particular video frame is considered as an excitement frame if its audio excitement or ZCR exceeds the threshold. The short time audio energy $E(n)$ and ZCR $Z(n)$ for frame n is computed as follows:

Short-time audio energy

$$E(n) = \frac{1}{V} \sum_{m=0}^{V-1} [x(m)w(n-m)]^2$$

where,

$$w(m) = \begin{cases} 1 \text{ if } \quad 0 \leq m \leq V-1 \\ 0 \text{ otherwise} \end{cases}$$

$x(m)$ is the discrete time audio signal, V is the number of audio samples corresponding to one video frame.

Short-time average zero-crossing rate
In discrete-time signals, a zero crossing is said to occur if successive samples have different signs. The short-time average zero-crossing rate $Z(n)$, as defined below, gives rough estimates of spectral properties of audio signals.

$$Z(n) = \frac{1}{2} \sum_{m=0}^{V-1} |sgn[x(m)] - sgn[x(m-1)]| w(n-m)$$

where,

$$sgn[x(m)] = \begin{cases} 1 \; x(m) \geq 0 \\ -1 \; x(m) < 0 \end{cases}$$

where, and $w(m)$ is a rectangular window. It is observed that audience cheering generally leads to high ZCR.

The strategy for the excitement detection is already explained in our previous work [10].

3.2 HMM Based Video Classification for Level 2 to 5

We have used HMM model to classify the trimmed video into one of the predefined classes. The class transition diagram is generated by training HMM through a number of sports video sequences and once trained, video shots can be classified into available classes by matching to the models of these classes. Our approach can be summarized as follows:

Step-1: Likelihood computations
Compute the likelihood $l_t(k)$ that the frame-t belongs to the class-k, based on the similarity of the features (such as color, motion etc.) of frame-t with those of class-k.

Step-2: Accumulated Likelihood Computation
Corresponding to the starting frame, the accumulated likelihood $L_t(k)$ for every class and the backtracking indices $A_t(k)$ for every class are initialized as follows:

$$L_1(k) = \alpha \; l_1(k)$$

and

$$A_1(k) = 0 \quad for \; k = 0, 1, 2, 3, ...N$$

where α is a multiplication constant.

For all subsequent frames, the accumulated likelihood and backtracking indices of every class is computed through a dynamic programming based optimum path search and is given by:

$$L_t(k) = \max_{1 \leq i \leq N}(L_{t-1}(i) + c(i,k)) + \alpha \; l_t(k)$$

and

$$A_t(k) = arg \max_{1 \leq i \leq N}(L_{t-1}(i) + c(i,k))$$

In the above equations, $c(i,k)$ indicates the transitional probability from class-i to class-k, determined through training. It is obvious from accumulated likelihood equation that higher value of multiplication factor α contributes to predominance of current likelihood over the accumulated ones.

Step-3: Frame-by-frame classification
Following step-2, the frames are classified individually, starting with the class C_t^* for the last frame of the sequence and continuing through a process of backtracking, as given below

$$C_T^* = arg \max_{1 \leq i \leq N}(L_T(i))$$

and

$$C_t^* = A_{t+1}(C_{t+1}^*),$$

$$where, \; t = T-1, T-2, ..., 1$$

3.3 Rule Based Activity Detection

To bridge the semantic gap, we have used the rule-based approach for level-4 for extracting semantic video concepts. The generic rule can be formed as follows

If{event A is followed by
post event activity ♯ 1 and/or post event activity ♯ 2
and/or· · · · · post event activity ♯ k}
Then {event A belongs to class i}
Else {event A belongs to class j }
Typical Examples of the rules for cricket video:
If {real time (L3: class-0) is followed by:
{Fielders' celebration (L3: class-1)} and/or {Batsman's departure (L3: class-2)}}
Then {action is wicket (L4: class-0)}
Else {action is hit (L4: class-1)}

Such many rules can be generated at different levels of hierarchical structure to extract the semantic concepts of the sports video.

4 Implementation and Results

We have tested our proposed approach using live recording of cricket and football video sequences. We sampled audio at a rate of 44.1 KHz. The performance of excitement detection was tested using the measure detection accuracy η_D, which is the ratio of number of excitement frames correctly detected, to the total number of actual excitement frames. Table 1 and Table 2 presents the detection accuracy for cricket video clip of 5:10 minutes and football video clip of 2:24 minutes respectively. For cricket video clip, we have extracted total 3109 video frames at the rate of 10 frames/second and for football video clip, we have extracted total 725 video frames at the rate of 5 video frames/second to increase computational speed. Fig 2 and Fig 3 show the graphs of audio energy and ZCR vs video frame number for cricket and football video respectively. We observed the average detection accuracy as 98.23% for cricket test video and 100% for football test video.

The overall performance of event classification is tested using the measure classification efficiency η_c, which is the ratio of the number of frames correctly classified to the total number of frames belonging to that particular class. Table 3

Table 1. Cricket video classification at Level 1 for various values of window size

Window size (sec)	Actual ♯ of excitement frames	♯ of excitement frames correctly detected	η_D %
5	1137	1094	96.22
10	1137	1110	97.63
20	1137	1132	99.56
40	1137	1129	99.30
50	1137	1119	98.42

Table 2. Football video classification at level 1 for window size of 10 seconds

Activity observed	Actual $\sharp$ of excitement frames	$\sharp$ of excitement frames correctly detected	η_D %
Foul	69 (1-68)	69	100
Goal miss	154 (175-329)	154	100
Free kick	140 (535-675)	140	100

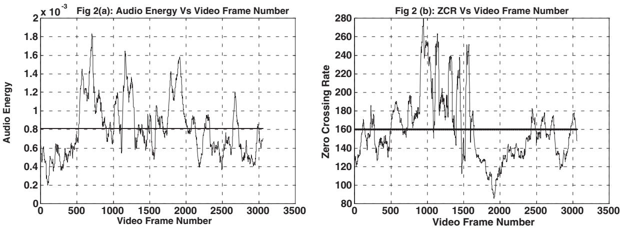

Fig. 2. (a) Audio Energy (b) ZCR Vs Video Frame Number for cricket video sequence

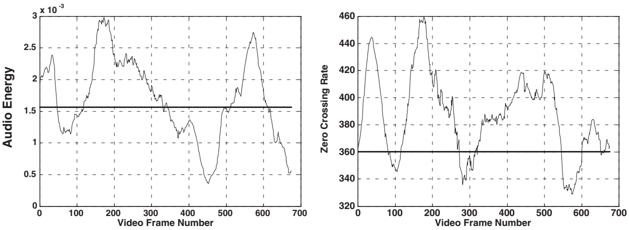

Fig. 3. (a) Audio Energy (b) ZCR Vs Video Frame Number for football video sequence

Table 3. Class Definitions of level-3

Class Number	Cricket	Football
0	Real time	Real time
1	Fielders' Celebration	Players' Celebration
2	Batsman's Departure	Players' Close-up
3	Replay	Replay

indicates our class definitions for level-3 for cricket and football video sequences. Table 4 and 5 represent the classification accuracy of cricket and football videos respectively.

Fig 4 shows boundary frames of the scenes of level-2 of cricket video, where we have used color as a likelihood function, since the color of ground is green and can be easily distinguished from spectator and commercial class. Fig 5 shows boundary frames of the scenes of level-3 of cricket video, where we have used color and motion as a likelihood function, since the color of ground is green and can be easily distinguished from fielders' celebration (where the dominance of

Table 4. Cricket video classification

Level	Beginning-end frame/ total frames/actual class	♯ of frames in observed class 0/1/2/3/..	η_c %	$\bar{\eta}_c$ %
2	525-894/370/0	314/38/18	84.86	94.28
	895-964/70/1	0/70/0	100	
	965-1661/697/2	0/14/683	97.99	
3	525-627/103/0	73/12/18/4	70.87	79.02
	628-733/106/1	2/83/21/8	78.30	
	734-790/57/2	1/1/54/1	94.74	
	791-894/104/3	12/8/11/74	71.15	
4	525-627/103/0	103/0	100	100
5	525-627/103/0.4	4/3/4/4/88	85.44	85.44

Table 5. Football video classification

Level	Beginning-end frame/ total frames/actual class	♯ of frames in observed class 0/1/2/3/..	η_c %	$\bar{\eta}_c$ %
2	175-329/154/0	154/0/0	100	100
	-/0/1	0/0/0	100	
	-/0/2	0/0/0	100	
3	175-198/24/0	19/2/2/1	79.16	87.30
	-/0/1	0/0/0/0	100	
	199-219/21/2	2/2/17/1	80.95	
	220-329/110/3	2/6/4/98	89.09	
4	175-198/24/1	0/24	100	100
5	175-198/24/1.2	3/2/17/2	70.80	70.80

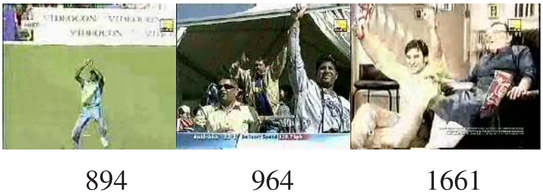

Fig. 4. Boundary frames of the scenes classified into class-0, class-1, and class-2 in level-2 of the cricket video

blue color is observed because our test video contains Indian fielders whose dress color is blue.) and batsman's departure (where the dominance of yellow color is observed because the color of Australian batsman's dress is yellow). We have also used motion as a likelihood to separate the real time action on the ground from the replays. Since the real time action is followed by fielders' gathering, our rule based classifier has declared the event in the cricket test video as a wicket.

Fig 6 shows boundary frames of the scenes of level-2, where we have used color as a likelihood function, since the color of ground is green and can be easily dis-

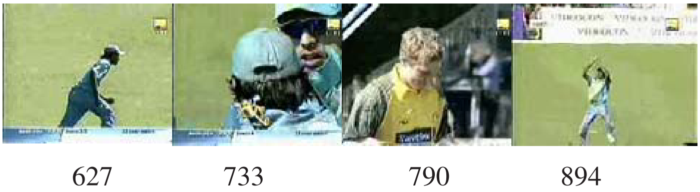

Fig. 5. Boundary frames of the scenes classified into class-0, class-1, class-2, and class-3 in level-3 of cricket video

Fig. 6. Boundary frames of the scenes classified into class-0, class-1 (no frame) and class-2 (no frame) in level 2 of football video

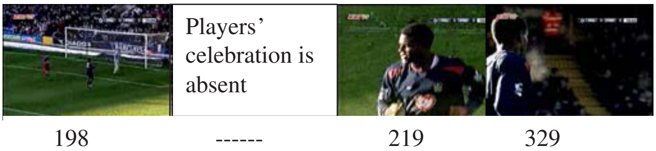

Fig. 7. Boundary frames of the scenes classified into class-0, class-1 (no frame), class-2 and class-3 in level 3 of football video

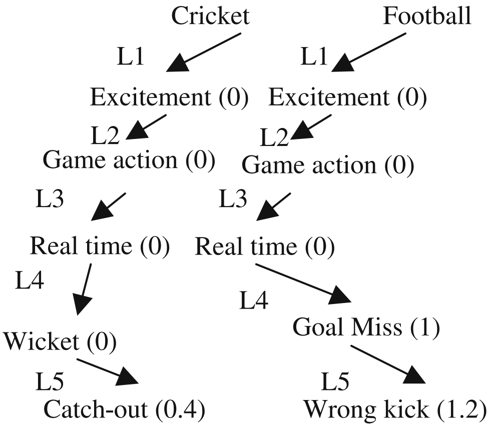

Fig. 8. Tree path followed by (a) cricket, and (b) football test video sequences

tinguished from spectator and commercial class. Fig 7 shows boundary frames of the scenes for level-3 for football video sequence where we have observed that the frames of class-1 are absent. This indicates that the players' celebration is absent. Hence our rule-based classifier has declared this activity as goal miss. Fig 8 shows the tree path followed by cricket and football test video sequences.

5 Conclusion and Future Work

In this paper, we have presented a generic hierarchical framework for sports video classification and successfully applied it to cricket and football. Integrating audio and video features for classifier not only reduces the cost of processing data drastically, but also increases the classifier accuracy significantly. The proposed modeling is readily applicable to media database management applications, where common operations such as indexing, retrieval, logging, annotation and highlights, etc can all benefit from the breakdown of a video into the smaller segments.

References

1. Kijak, E., Gravier, G., Gros, P., Oisel, L., Bimbot, F.: Hmm based structuring of tennis videos using visual and audio cues. in Proc. of Int. Conf. on Multimedia and Expo **3** (2003) 309–312
2. Assfalg, J., Bertini, M., Bimbo, A.D., Nunziati, W., Pala, P.: Detection and recognition of football highlights using hmm. in 9th Int. Conf. on Electronics, Circuits and Systems **3** (2002) 1059–1062
3. Duan, L., Xu, M., Tian, Q., Xu, C.: Nonparametric color characterisation using mean shift. Proc. of 11^{th} ACM Int. Conf. on Multimedia (2003) 243–246
4. Xie, L., Chang, S.F., Divakaran, A., Sun, H.: Structure analysis of soccer video with hidden markov models. in Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (2002)
5. Baoxin, L., Pan, H., Sezan, I.: A general framework for sports video summarization with its application to soccer. in Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing **3** (2003) 169–172
6. Hanjalic, A.: Generic approach to highlights extraction from a sports video. in Proc. of IEEE Int. Conf. on Image Processing **1** (2003) 1–4
7. Hsu, W., Kennedy, L., Huang, C.W., Chang, S.F., Lin, C.Y., Iyengar, G.: News video story segmentation using fusion of multi-level multi-modal features in trecvid 2003. in Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing **3** (2004) 645–648
8. Chaisorn, L., Chua, T.S., Lee, C.H.: The segmentation of news video into story units. in Proc. of Int. Conf. on Multimedia and Expo **1** (2002) 73–76
9. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. in Proc. of IEEE **77** (1989)
10. M.H.Kolekar, Sengupta, S.: Hierarchical structure for audio-video based semantic classification of sports video sequences. in Proc of SPIE Int. Conf. on Visual Communications and Image Processing, Beijing, China **5960** (2005) 401–409