EVENT DRIVEN CONTENT STRUCTURE ANALYSIS OF TENNIS VIDEO

Vyacheslav Parshin, Liming Chen

LIRIS, Ecole Centrale de Lyon
36, Avenue Guy de Collongue, 69131 Ecully, France
{vyacheslav.parshin, liming.chen}@ec-lyon.fr

ABSTRACT

An approach to automatic tennis video segmentation is proposed. The aim is temporal decomposition of a tennis match according to its hierarchical semantic content structure which could be used to organize an efficient content based access. The approach relies on some particular characteristics that facilitate to convey semantic information to a viewer. In this work they are specific views and score boards which are automatically detected and serve as events driving the content parsing. We propose quite a general framework which can be considered as a kind of a final state machine whose states relate to content units. Advantage of our approach is in its expressiveness and low computational complexity. Experimental evaluations on ground-truth video were made that showed quite high segmentation accuracy.

1. INTRODUCTION

Sports video has usually a well-defined temporal content structure which could be used to efficiently organize a content-based access that allows for such functions as browsing and searching, as well as filtering interesting segments to make compact summaries. According to the logic of a tennis match, it can be naturally represented in a hierarchical manner as a sequence of sets that in their turn are decomposed into games etc. In this work we propose an approach to automatic tennis video parsing that yields a temporal decomposition of a given video into such a hierarchical content structure.

To detect regular content units of tennis video we rely on some particular characteristics and production rules that are typically employed to convey semantic information to a viewer. Like a lot of other sports games, a tennis match is usually shot by a number of fixed cameras that yield unique views during each segment. For example, a serve typically begins with switching of the camera into the global court view (see figure 1). Since a match occurs in a specific playground, this view can be detected based on its unique characteristics (we employ its color homogeneity property). In order to constantly keep the audience informed about the current game state, score or statistics boards are regularly inserted into the broadcast according to the game progress. In our content parsing technique these inserts are detected and used as indicators of transitions between semantic segments.

Figure 1. Global court view samples.

The works concerned with sport video content retrieval problem usually aim at detection of only some specific semantic events or scenes from low-level visual and audio features using domain-specific models or pattern recognition techniques [1]-[3]. The aim of our work, however, is to develop a quite general framework that allows us to combine different semantic-level events for the purpose of full content parsing, rather than elaborate domain-specific detectors.

2. SEGMENTATION ENGINE

We represent the tennis content structure hierarchically as a sequence of nested temporal segments. Different structures can be proposed depending on the needs of a user. In this work we use a configuration presented in figure 2. It shows segment types allowed at each semantic level; segments of a higher level can comprise segments of several types in the lower level. This configuration corresponds to the logical structure of a tennis match.

The segmentation is performed in two stages. First, intermediate semantic events are detected from an input video; in this work they are score boards and a global court view. The beginning and the end of these events are mixed so as to form a sequence of time ordered instantaneous events which is used at the stage of content generation. The multilevel content structure of video is generated recursively, beginning at the highest semantic

level. Each semantic segment is detected using its proper event or, in more complicated case, the pattern of corresponding events combined by temporal relations.

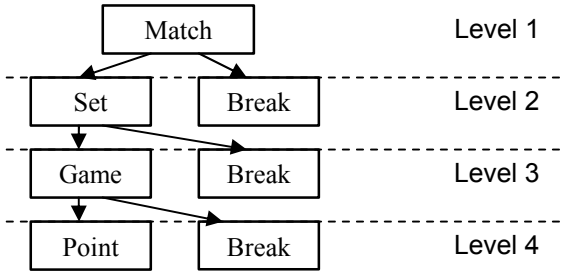

Figure 2. Tennis video content structure.

The content table is generated by a kind of a state machine whose states correspond to the appropriate semantic segments. The parsing is driven by its grammar that imposes state transition constraints, event pattern detectors that control the transition from one state to another and time adjustment events that indicate the precise position of segment boundaries. Parsing rules developed for the content structure of the figure 2 are given in tables 1a, 1b and 1c. The "transition pattern" and the "time adjustment event" correspond to the transition to a state listed in the first column of the tables. In the general case it is supposed that the initial segment of a given video is unknown. That is why the state machine starts from initial undefined state at the second semantic level. For the lower semantic levels the initial state is chosen according to table 1a-1b. Our recursive parsing algorithm for a given semantic level is the following:

- For each event extracted in the time order do:
  - If the event together with the previous events corresponding to the current machine state constitute allowable transition pattern, then:
    - Find the appropriate time adjustment event for the current state and correct its beginning time accordingly.
    - If the semantic segment corresponding to the previous machine state has to be further decomposed into the segments of the lower level, then initialize the current state for that level accordingly and perform the parsing recursion for that segment.
    - Go to the next machine state according to the detected transition pattern.
- If the resting semantic segment corresponding to the last machine state has to be further decomposed into the segments of the lower level, perform the parsing recursion for that segment.

The rules of table 1a imply that a tennis set begins with the first serve. Therefore it is detected at the transition to the global court view which normally accompanies a serve. A unique score/statistics board is usually inserted a little time after the end of a set which is defined as the end of the last rally. Hence, we detect a break at the beginning of the corresponding score board and then shift its beginning to the transition from the last global court view which normally shows a rally. The rules of table 1b and 1c are derived in a similar way. We use a pattern instead of a single event to detect a transition from a game to a break to exclude false transition in the case when the producer inserts a score board during the first global view of a game. The "during" time relation is defined according to Allen's notation [4].

State	Allowable next states	Initial state of the sublevel	Transition pattern	Time adjustment event
Initial undefined	Set	-	-	-
Set	Break	Game	Transition to a global view	-
Break	Set	-	Appearance of a set score board	End of the last global view corresponding to the previous state

Table 1a. Parsing rules for semantic level 2 (of tennis sets).

State	Allowable next states	Initial state of the sublevel	Transition pattern	Time adjustment event
Game	Break	Point	Transition to a global view	-
Break	Game	-	Appearance of a game score board not during a global view	End of the last global view corresponding to the previous state

Table 1b. Parsing rules for semantic level 3 (of tennis games).

State	Allowable next states	Initial state of the sublevel	Transition pattern	Time adjustment event
Point	Break	-	Transition to a global view	-
Break	Point	-	Appearance of a point score board	End of the last global view corresponding to the previous state

Table 1c. Parsing rules for semantic level 4 (of tennis points).

3. EVENTS DETECTION

3.1. Global court view

Tennis video like a lot of other types of sports video is usually shot by a fixed number of cameras that give unique views for game segments. A transition from one such view to another is sometimes an important indicator of semantic scene change. In tennis video a transition to the global court view that shows the whole field area with the players commonly signifies that a rally begins. When the rally finishes, a transition to another view such as a player close-up or the audience usually happens. Thus, court view recognition is important for rally scenes detection.

The first step in the detection of a specific view is segmentation of the video into views taken by a single camera or, in the other words, segmentation into shots. Color histogram difference between consecutive frames is applied in order to detect shot transitions. We use 64-bins histograms for each 3 components of the RGB-color space and concatenate them into one 192-dimensional vector. To be able to detect both abrupt and gradual transitions in the video, the twin-threshold method [5] is adopted.

The color distribution of global court view shots does not change much during the tennis match. This allows us to detect them based on their comparison with a sample frame of the court view which is selected manually at the learning stage. A shot is recognized as a global court view if it is close enough to the appropriate sample view. Two sample frames selected at the learning stage of experimental evaluations are shown in figure 1. The learning sample is selected only once for a game or a series of games played at the same court (e.g. during the same championship).

In tennis video there are usually several types of shots that contain a big part of the tennis field at the background and, thus, resemble much the global court views such as player close-ups. However, the court views usually take a longer part of the tennis video. Hence, we enhance the robustness of the court view detection by grouping the shots into similarity clusters and, then, rejecting rare clusters.

3.2. Score board detection

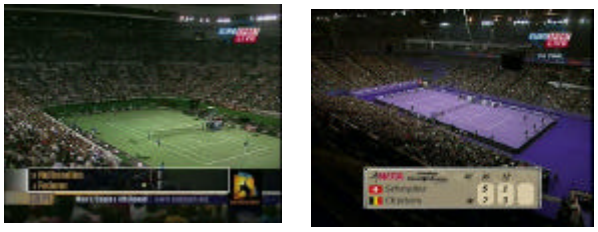

Figure 3. Score boards inserted btw. tennis games.

As reflecting the state of the game, score boards could provide useful information for tennis video parsing into its logical structure. Since these boards are inserted regularly according to the game progress, the mere facts of their appearance/disappearance can be used as reliable indicators of the semantic segment boundaries. Moreover, they present important information about the game and, hence, we can choose the appropriate frames as the key frames of the corresponding semantic units and thus provide convenient visual interface for browsing through the content table.

The same tennis video usually has several types of score boards that can be used to separate the segments at different levels of the semantic hierarchy. Score boards of the same type have the fixed positions on the screen and similar color bitmaps. The only difference between them lies in their textual content, the horizontal size (which is changed so as to hold all required data) and somewhat in their color (caused by the partial transparency). Several sample frames which contain score boards along with their bounding rectangle are shown in figure 3. We detect score boards, if we find horizontal lines of enough length placed near their upper and bottom borders. The Hough transform is applied to edge points in order to detect the lines. The positions of the score boards borders are given manually during the learning – a user selects from sample tennis video the frames that contain required score tables and picks out their bounding rectangle.

4. EXPERIMENTAL EVALUATIONS

The performance of our parsing system was experimentally evaluated on three tennis video records

captured from Eurosport satellite channel. One of them shows an excerpt of a tennis match of Australia Open (AO) 2003 championship, two others represent fragments of two matches of WTA tournament. The former lasts about 51 minutes, the rest two – 8.5 and 10 minutes. The two tournaments have different score board configuration and color distribution of the court. So, we extracted two sets of learning samples for the events detectors.

The results of segmentation accuracy evaluations based on comparisons with manually labeled data are presented in table 2. The automatic segmentation was performed using parsing rules of table 1a-1c. Semantic levels 3 and 4 (see figure 2) were treated separately; level 2 was not considered as there are too few set segments in the ground truth. Recall r and precision p are computed as:

$$r = \frac{n_c}{n_c + n_{miss}}, \quad p = \frac{n_c}{n_c + n_{f.a.}}, \qquad (1)$$

where n_c, n_{miss} and $n_{f.a.}$ are the number of correct, missed and false alarm boundaries respectively. Detected segment boundary was considered as correct if it coincided with corresponding manual one within ambiguity of 1 sec. Otherwise it was considered as a false alarm. A manual boundary was considered as missed one if it did not coincided with corresponding automatically detected ones within the same ambiguity of 1 sec. In order to reduce the influence of "edge effects" on the segmentation evaluations results, the first and the last segments of the lowest semantic level were cut off by half from the comparison intervals for each video record. The results of classification accuracy evaluations total for both the tournaments are given in table 3 (recall and precision are computed using expression (1) where the term n stands for the duration of the corresponding segment part).

As for processing time, our parsing technique is quite fast provided that the events are already extracted and takes less than 1 second for a usual tennis match on modern personal computers. This is because the computational complexity is approximately proportional to the number of events and the number of semantic levels. The major computational power is required to decompress the video and detect the relevant events. On our Intel Pentium 4 1.8 GHz computer this task is performed nearly in real time for MPEG1 coded video, though we did not make a lot of optimizations.

The most of the segmentation errors are caused by unreliability of event detectors. High rate of false score boards result in relatively low precision of segmentation on games and breaks for AO tournament. It is caused by resemblance of the score board, which is a true indicator of the segment transitions, to a statistics board which was inserted in any place during games. One of the sources of the errors at semantic level 4 is a high false alarm rate for global court views which is caused by confusions with replay shots (they shift the transition between a point and a break). There are only few segmentation errors at the semantic level 4 for AO tournament that steam from the parsing rules. They are caused by the fact that sometimes the producer forget to show a score board or insert it after the first serve of a point.

Tournament	Semantic level	Recall	Precision	F1
AO	3	0.84	0.62	0.71
	4	0.82	0.91	0.86
WTA	3	1	0.83	0.91
	4	0.94	0.98	0.96
AO+WTA	3	0.90	0.68	0.78
	4	0.86	0.93	0.89

Table 2. Segmentation results.

Semantic Level	Segment	Recall	Precision	F1
3	Game	0.97	0.99	0.98
	Break	0.97	0.91	0.94
4	Point	0.83	0.98	0.90
	Break	0.97	0.89	0.93

Table 3. Classification results total for both the tournaments.

5. CONCLUSIONS

Quite a general framework is proposed for hierarchical content parsing of tennis video. It is based on some particular characteristics and production rules that are typically employed to convey semantic information to a viewer, such as specific views and score boards. The advantage of our approach is in its expressiveness and low computational complexity. Moreover, the experimental evaluations showed quite high segmentation accuracy on ground-truth data.

6. REFERENCES

[1] Di Zhong, Shih-Fu Chang, "Structure Analysis of Sports Video Using Domain Models", *Proc. IEEE ICME'01*, Japan, pp.182-185, August 2001.

[2] Rozenn Dahyot, Anil Kokaram, Niall Rea and Hugh Denman, "Joint audio visual retrieval for tennis broadcast", *Proc. IEEE ICASSP*, Hong Kong, April 2003.

[3] Z.Zivkovic, F. Heijden, M.Petkovic, W.Jonker, "Image processing and feature extraction for recognizing strokes in tennis game videos", *Proc. of 7th Annual Conference of the Advanced School for Computing and Imaging,* the Netherlands, pp.512-516, June 2001.

[4] J.F.Allen, "Maintaining knowledge about temporal intervals", *Communications of the ACM 26*, pp. 832-843, 1983.

[5] Dong Zhang, Wei Qi, Hong Jiang Zhang, "A New Shot Boundary Detection Algorithm", *IEEE Pacific Rim Conference on Multimedia*, pp.63-70, 2001.