DETECTION OF SLOW-MOTION REPLAY SEGMENTS IN SPORTS VIDEO FOR HIGHLIGHTS GENERATION

H. Pan *P. van Beek* *M. I. Sezan*

Electrical & Computer Engineering
University of Illinois
Urbana, IL 61802
haopan@uiuc.edu

Sharp Laboratories of America
5750 NW Pacific Rim Blvd.
Camas, WA 98607
pvanbeek@sharplabs.com sezan@sharplabs.com

ABSTRACT

In this paper, we present a novel method for generating sports video summary highlights. Specifically, our method localizes semantically important events in sport programs by detecting slow motion replays of these events, and then generates highlights of these events at multiple levels. In our method, a hidden Markov model (HMM) is used to model slow motion replays, and an inference algorithm is introduced which computes the probability of a slow motion replay segment, and localizes the boundaries of the segment as well. An effective new feature is used in our HMM, based on a moving measure of the number of zero-crossings and the amplitudes of variations over time of video field differences. Furthermore, the method is capable of filtering out slow motion play segments in commercials. As compared with existing methods for video event detection, our method is more generic (i.e., domain independent), and has the ability to capture inherently important events.

1. INTRODUCTION

With the development of high-speed Internet, high-capacity storage, and high-ratio compression standards such as MPEG -1, -2 and -4, people are quickly drowning in a growing amount of available video information. Therefore, automatic detection of semantically important events in video and further summarization of video to help indexing, browsing and consuming the video has become increasingly important.

Many approaches towards automatic event detection and summarization in sports programs have been reported in literature, e.g. [1-5]. However, most methods are developed for particular sports, specific edit effects, or specific environments only, resulting in domain specific approaches. For example, some require the events to take place in sites under surveillance [1], some are restricted to football games [2], some are restricted to baseball [3], to basketball [4], or to soccer [5].

In this paper, we propose an entirely novel, more generic method. Based on the observation that in sports programs, important events are often replayed in slow motion immediately after they occur, our method detects slow motion replay segments to localize semantically important events and then further summarize sports programs. A clear advantage of our method over the existing domain-specific methods is that this strategy is generic in nature, and is therefore applicable to any sports, and in fact any other kinds of video programming. Furthermore, information about replay segments can also be used in combination with other types of information obtained by the existing methods.

While we propose to use localization of SLO-MOs for event detection and summarization, Kobla et al. [6-7] detect SLO-MOs in sports programs as a feature for sports/non-sports video classification. The method reported in [6-7] is block-based, which uses motion vectors in the MPEG-1 domain, while our method is pixel-based. No attempt is made in [6-7] to localize the boundaries of slow motion replay segments, or to distinguish SLO-MOs in commercials from the relevant SLO-MOs. Overall, we believe our method is more generic and accurate regarding localization of slow motion replay segment boundaries.

This paper is organized as follows. In Section 2 we discuss the structure of slow-motion replay segments in sports programs. In Section 3, we discuss the proposed method, which has three major parts: (1) detecting SLO-MOs; (2) distinguishing between SLO-MOs containing program events and SLO-MOs that may be in the commercials included in the program; (3) using SLO-MOs in generating program highlights at different durations. Finally, we present experimental results, followed by the conclusions.

2. SLOW MOTION REPLAY SEGMENTS IN SPORTS VIDEO

Figure 1 contains a simplified diagram of the structure of slow motion replay segments in sports video. The action shots containing the important event are often followed by other shots before the slow motion replay segment, which itself usually contains editing effects at the front and at the end. In this paper, we assume that sports video programs are in the format of 60 fields/second interlaced NTSC.

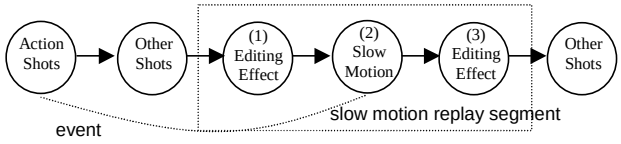


Figure 1. The structure of slow motion replay segments.

We further classify the fields in slow motion replay segments (SLO-MOs) as follows.

(1) & (3) *Editing effects fields* (in & out), which mark the start and end of SLO-MOs, occupy less than 10% of SLO-MOs. Any gradual transitions, such as fade in & out, cross/additive-dissolve and wipes, can be used in *editing effects*.

(2) *Slow motion fields*, which form the visual slow motion effect, occupy more than 90% of SLO-MOs. The slow motion effect is usually attained by one of two methods: if the video was recorded by standard cameras, certain fields are simply repeated; if the video was recorded by 3-time high-speed super

motion cameras, fields are played out at the normal playing speed. In the latter method, the view effect is fixed at exactly 3 times as slow as the normal speed, if all the recorded fields were played back. However, the play back speed during a slow motion is usually controlled manually by hand, and frequently, some of the fields may be repeated or dropped to make the play speed slower or faster for a better visual effect. Thus, both methods are characterized by field repetition and/or field drop. The method of standard camera + field repetition is much more widely used, because it is easy and cheap to implement, and the visual effect is satisfactory for most sports video programs.

Besides *editing effects* and *slow motion replay* fields, there are two other types of fields, *still* fields and *normal motion replay* fields, which are not shown in Figure 1 for simplicity. These two types of fields do not always exist in SLO-MOs, but they may occur between *editing effects* and *slow motion replay* fields, or between *slow motion replay* fields.

Finally, it is important to point out that SLO-MOs also prevail in commercials. Thus, for video event detection and highlights generation, it is critical to distinguish SLO-MOs in sports games from ones in commercials.

3. THE PROPOSED METHOD

The proposed method has three components: a detector of SLO-MOs, which detects SLO-MOs and localizes their boundaries; a commercial/non-commercial filter, which filters out SLO-MOs in commercials; a summary generator, which generates program highlights at different durations from filtered SLO-MOs.

3.1 Slow motion detection

We use a hidden Markov model (HMM) [9] to model the relations of the five types of fields in SLO-MOs, to localize the boundaries, and to calculate the probability of every SLO-MO candidate.

The structure of SLO-MOs described in Section 2 can be well modeled by an HMM. However, a conventional HMM using the Viterbi algorithm only computes the probability of a fixed-length input sequence, and thus a conventional HMM doesn't have the ability to localize a portion in a sequence that best fits the HMM, which is our requirement. We address this problem in the following paragraphs, where we introduce a specific structure of the HMM and an inference algorithm, capable of localizing boundaries of SLO-MOs.

3.1.1. HMM structure

The HMM, shown in Figure 2, is built to model a half of SLO-MOs, starting from *slow motion* fields in a SLO-MO, either going forward or backward, and ending at *normal play* fields before or after the SLO-MO. The HMM has five states: *(0) slow motion*, *(1) still*, *(2) normal replay*, *(3) edit effect in/out*, and *(4) normal play*. States (0)-(3) respectively correspond to all the four types of fields in a SLO-MO, described in Section 2. Note that we introduce state (4), *normal play*, which is mapped to the fields immediately outside a SLO-MO. State (4), associating with the inference algorithm, plays a crucial role in localizing the boundaries.

The inference algorithm of this HMM, based on the Viterbi algorithm, is as follows.

1. Use a simple normalization + threshold method to pinpoint with high probability a single field inside in a SLO-MO;
2. Use that field, defined as the origin, as the starting point for a forward and a backward pass, each of length L fields, where L is long enough to contain the boundaries;
3. Feed the L "forward-pass" fields into the HMM and run the Viterbi algorithm to determine the optimal state sequence. The first field that reaches the hidden state (4) is the boundary that ends the SLO-MO;
4. Feed the L "backward-pass" fields into the HMM and run the Viterbi algorithm to determine the optimal state sequence (note the backward sequence is treated in time reversed order). The first field that reaches the hidden state (4) is the boundary that starts the SLO-MO.

By introducing an extra hidden state and two-pass inference algorithm starting from the middle of a SLO-MO, this HMM is capable of localizing the boundaries. In this paper, we choose L=800. Because SLO-MOs are usually shorter than 20 seconds, L=800 is long enough to contain the starting and ending boundaries (given the field rate of 60 fields/second).

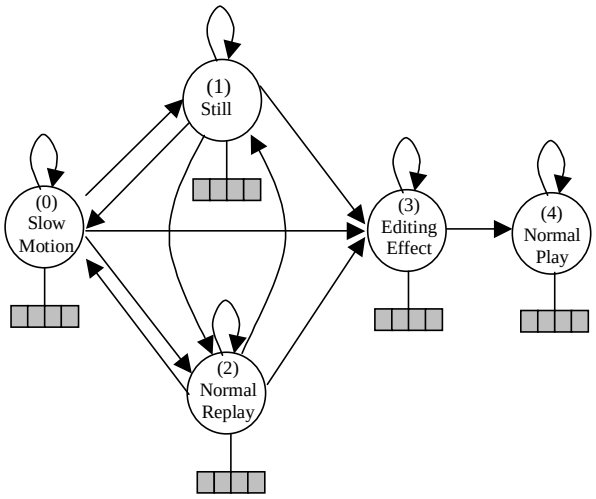


Figure 2. The structure of the Hidden Markov Model.

3.1.2. HMM features

Four features are used in the HMM, three of which are calculated from the pixel-wise mean square difference of the intensity of every two subsequent fields, which is denoted by $D(t)$, and one of which is computed from the RGB color histogram of each field.

The three features based on $D(t)$ are: (a) a measure of zero-crossings in a sliding window over $D(t)$ along the time axis; (b) the lowest value of $D(t)$ in the sliding window; and (c) the differences of every two adjacent values of $D(t)$. The sliding window is S fields long and moves forward 1 field each time. These three features describe the *still*, *normal motion replay*, and *slow motion* fields.

Slow motion fields are generated by field repetition/drop, and field repetition/drop cause frequent and strong fluctuations in $D(t)$, which can be measured by a zero-crossing measure, $p_{zc}(t)$. This measure is defined in the following two steps.

First, we define the number of zero-crossings in a window of length S fields as

$$zc(t,\theta) = \sum_{s=1}^{S-1} trld\left(D(t-s) - \overline{D}(t), D(t-s-1) - \overline{D}(t), \theta\right)$$

where $\overline{D}(t)$ is the average of $D(t)$ over a sliding window at time t,

$$trld(x,y,\theta) = \begin{cases} 1 & if\ x \ge \theta\ \&\ y \le -\theta\ or\ x \le -\theta\ \&\ y \ge \theta \\ 0 & else \end{cases},$$

and θ is a threshold on amplitudes of fluctuations in D(t).
Next, we quantize the effect of θ on $zc(t,\theta)$. By introducing Θ, a set of ascendant thresholds θ_i indexed by $i=1,2,\cdots,I$, we define

$$p_{zc}(t) = \begin{cases} \arg\max_i \{\theta_i \mid zc(t,\theta_i) > \beta, \theta_i \in \Theta\} & if\ zc(t,\theta_1) > \beta \\ 0 & else \end{cases}$$

where β is a threshold on $zc(t,\theta_i)$. Note that once $zc(t,\theta_i)$ passes β, $p_{zc}(t)$ is dependent on the amplitudes of fluctuations in $D(t)$.
The zero-crossing measure $p_{zc}(t)$ is illustrated in Figure 3. This zero-crossing measure takes into account both the frequency and amplitude of the fluctuations.

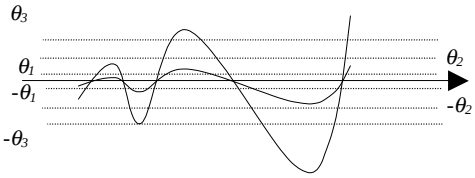


Figure 3. The zero-crossing measure. Note that the numbers of zero-crossings of the two curves are the same if the difference in amplitudes of fluctuations is not taken into account. On the other hand, setting $\beta=2$, the low-amplitude curve yields $zc(t,\theta_1)=3$, $zc(t,\theta_2)=0$ and $zc(t,\theta_3)=0$. Because only $zc(t,\theta_1)>\beta$, $p_{zc}(t)=1$; The high-amplitude curve yields $zc(t,\theta_1)=5$, $zc(t,\theta_2)=4$ and $zc(t,\theta_3)=3$. Because all the three are bigger than β, we select the biggest θ_i and assign the corresponding i to $p_{zc}(t)$. Therefore, $p_{zc}(t)=3$.

The fourth feature, based on the color histogram, is for capturing the gradual transitions in *editing effects*. There are many papers addressing this problem. We have adopted the method described in [8].
All the four features are normalized and quantized to 16 levels before they are used by the HMM.

3.2 Commercial/non-commercial filtering

As discussed in Section 2, there may also be SLO-MOs in the commercials that are included in the sports programming, as is the usual case in TV broadcasts. Thus, we need a commercial/non-commercial filter to distinguish between SLO-MOs in commercials and in the actual program.
The principle of the commercial/non-commercial filter is that the average color histogram of a commercial SLO-MO is quite different from the average color histogram of a segment in the game while all the segments in the game have similar histograms. Thus, once we have identified segments that are part of the game (not necessary SLO-MO segments), we can use them as references and filter out commercial SLO-MOs, by comparing the distances of color histogram of SLO-MOs with the average color histogram of the references. Because each commercial is less than 2 minutes long, and the interval between two commercial interruptions is longer than 5 minutes, the two positions that are two minutes before and after the SLO-MOs must be non-commercial, and thus serve as the references.

3.3 Video summary highlights generation

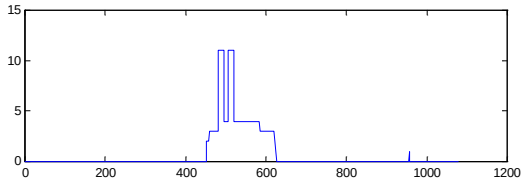
Based on information about the location of SLO-MOs, we generate highlight summaries of the video program. The resulting highlights may include multiple levels of highlights with varying detail, ranging from short to longer highlights, as follows.

1. Concatenation of all non-commercial SLO-MOs. The resulting highlight provides the most compact summary of the program (and it does not contain any commercials).
2. Concatenation of expanded non-commercial SLO-MOs. Expansion is performed by adding t_1 and t_2 seconds to the beginning and end of each SLO-MO.
3. Same as level 2, but the expansion time intervals are chosen as a function of the statistics of the corresponding SLO-MO. In one possible implementation, the value of t_1 is set proportional to the actual length of the event depicted in the SLO-MO. The value of the proportionality factor k is determined by the length of the desired summary. To avoid overlaps between segments, a simple check mechanism can be introduced through which the value of k is controlled adaptively.

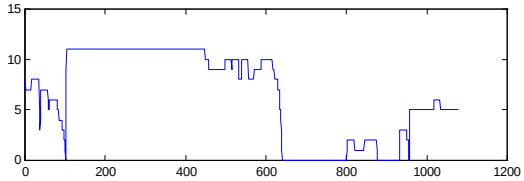
4. EXPERIMENTAL RESULTS

Our experimental data were captured from 4:2:2 YUV NTSC D-1 tapes, which have a 720x243 resolution for each field and a 60Hz field rate. We down-sampled the resolution to 360x240 to reduce the computational cost. We captured 10 different video clips in total, about 20 Gigabytes of data, with a duration of 25 minutes long. The clips cover five different types of games. Among them, there are two 10-minute clips: one basketball and one football, and 8 much shorter clips (usually shorter than 1 minute): one auto racing, three basketball, one boxing, one football and one soccer. We used 5 shorter clips to train the HMM, and use all clips as the testing data (We didn't observe different performance between the training and other data during the test).
There are a total of 91,204 fields in all clips, and 23,718 fields of them are in total 15 non-commercial SLO-MOs. The SLO-MOs detector detected all the 15 non-commercial SLO-MOs and additional 8 commercial SLO-MOs in the clips. Thus, the detection of non-commercial SLO-MOs has a success rate of 100%. The commercial/non-commercial filter filtered out 7 commercial SLO-MOs while preserving all the non-commercial SLO-MOs and one commercial SLO-MO. Therefore, the commercial/non-commercial filter has a success rate of 95.7%. Regarding boundary localization accuracy, we have used two measurements. The first one is defined as the ratio of the miss-detected fields over the total number of fields in the SLO-MOs. The second one is defined as the ratio of miss-detected fields over the total number of fields in the clip. The localization performance according to the former definition was 12.81%, while according to the latter it was 3.33%.
We have used the two 10-minute clips to generate multi-level summary highlights (the other seven clips are too short to generate a highlight other than the level-1 highlights discussed in Section 3.3, which only contain SLO-MOs). The highlight summaries of the two 10-minute clips are generated at 3 different levels, ranging from around 100 seconds to 4 minutes long. The authors believe that the generated highlights summarize the most important events in these clips very well. After we further added gradual transitions between different scenes, the highlights are visually pleasing overall as well.
The normalized and quantized values of the four features and the final hidden states of the HMM for a short soccer clip (1060 fields long) are shown in Figure 4.
While detection of SLO-MOs is reliable, localization of boundaries of SLO-MOs is not perfect. Sometimes, scene-cuts occur within a SLO-MO without editing effects. If there are *normal motion play* fields in such a SLO-MO, it is difficult to

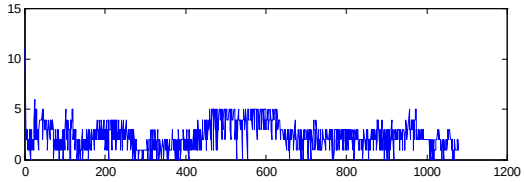
distinguish the end of a SLO-MO from a scene-cut within a SLO-MO, resulting in a significant error. Probably, additional features are necessary (for instance, audio features) to solve this problem.



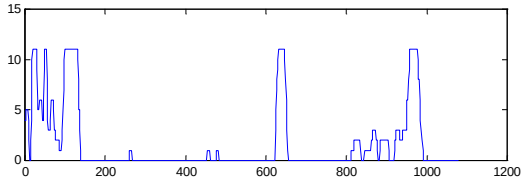
(a) The feature from $p_{zc}(t)$.



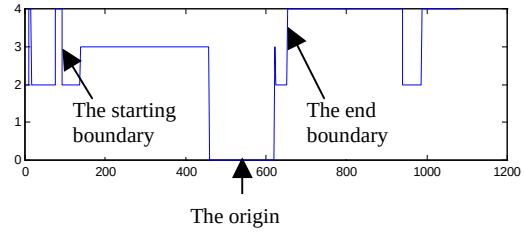
(b) The feature from the lowest value in a sliding-window.



(c) The feature from the differences of every two adjacent values of $D(t)$.



(d) The feature from the color histogram.



(e) The optimal hidden states ot the HMM, the starting and ending boundary points and the origin point.

Figure 4. The four features of the HMM and the optimal hidden states obtained by the Viterbi algorithm in a short soccer clip.

5. CONCLUSIONS & DISCUSSION

In this paper, we propose a new method for automatic event detection and summarization of sports programming. The method is based on the notion that slow motion replay segments are important clues to localizing semantically important events. By detecting slow motion replay segments, the method finds the locations of inherently important events in lengthy programs, and further generates multi-level summary highlights. In our paper, the structure of slow motion replay segments is analyzed, and a Hidden Markov Model is proposed which matches the structure of slow motion sequences. An inference algorithm is used to detect the boundaries of slow motion replay segments, and we have developed four features for our HMM framework, including a new zero-crossing measure. We also proposed a solution for filtering out slow motion segments that are part of commercials. Finally, a summary of program highlights is generated by concatenating the (non-commercial) slow motion replay segments with segments immediately before and after the detected segments.

Our method can be utilized at either the program provider or the consumer side to provide highlights of sports events that may vary in duration according to personal preferences, usage conditions, and requirements on system resources.

Currently, all four features used in the HMM framework are derived from the video signal. Other features, based on the audio signal, can be introduced into the framework readily, to enhance the performance and robustness of our algorithm.

6. REFERENCES

[1] J.D. Courtney, "Automatic video indexing via object motion analysis," *Pattern Recognition*, vol. 30, no. 4, pp. 607-626, 1997.

[2] S.S. Intille, "Tracking using a local closed-world assumption: Tracking in the football domain," *Proc. SPIE Storage and Retrieval for Image and Video Databases*, pp. 216-227, 1997.

[3] T. Kawashima, K. Tateyama, T. Iijima, and Y. Aoki, "Indexing of baseball telecast for content-based video retrieval," *Proc. IEEE Int. Conf. Image Processing*, pp. 871-875, 1998.

[4] D. Saur, Y.-P. Tan, S.R. Kularni, and P. J. Ramadge, "Automated analysis and annotation of basketball video," *Proc. SPIE Storage and Retrieval for Image and Video Databases*, pp. 176-187, 1997.

[5] D. Yow, B. L. Yeo, M. Yeung, and G. Liu, "Analysis and presentation of soccer highlights from digital video," *Proc. Asian Conf. Computer Vision*, 1995.

[6] V. Kobla, and D.S. Doermann, "Detection of slow-motion replays for identifying sports videos," *Proceedings of IEEE Third Workshop on Multimedia Signal Processing*, pp. 135-140, 1999.

[7] V. Kobla, D. DeMenthon, and D. Doermann, "Identification of sports videos using replay, text, and camera motion features, " *Proc. of the SPIE Conference on Storage and Retrieval for Media Databases*, Vol. 3972, pp. 332-343, Jan, 2000.

[8] S. Golin, "New Metric to Detect Wipes and Other Gradual Transition in Video," *Proc. of IS&T/SPIE Conference on Visual Communications and Image Processing*, 1999.

[9] L. Rabiner and B.-H. Huang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.