ONLINE END DETECTION FOR LIVE-BROADCAST SPORTS TV PROGRAMS

Hao-Da HUANG
Key Lab of Computer Science, Chinese Academy of Sciences
hdhuang@ios.ac.cn

Xian-Sheng HUA, Shipeng LI, Hong-Jiang ZHANG
Microsoft Research Asia
{xshua; spli; hjzhang}@microsoft.com

ABSTRACT

In this paper, a method for automatically detecting the end of lively broadcasted sports programs is proposed, which enables users to record the full TV programs when they run over time. Taking advantage of the property of relatively high content consistency within sports programs, this method is based on checking the break point of this content consistency. A scalable video segment similarity measure is proposed to measure content consistency of video segments in different similarity levels. Based on this measure, a two-round end detection scheme is applied, in which the first round, *Candidate End Point Finding*, is able to find a coarse candidate end point, and the second round, *End Point Refining*, is able to find a more accurate end. Experiments show that the proposed end detection scheme is able to detect the real ends with high accuracy.

1. INTRODUCTION

Digital Video Recorder (DVR) is widely used to record TV shows to hard disks, which makes it much easier for users to freely watch their favorite TV shows, as well as conveniently to use the recordings in further applications. With DVR, users are able to automatically record their favorite TV programs according to Electronic Program Guide (EPG), in which the start and end time of the program can be obtained online through a network. Generally most TV programs have fixed play durations as have indicated in EPG, thus they can be fully recorded by typical DVRs. However, live programs such as sports game and award shows often run over time, which may end even more than one or more hours after the scheduled end time in EPG. Though public TV stations in some countries use Vertical Blanking Interval (VBI) to transmit online information (say, Video Program System (VPS) in European countries) thus the DVRs can detect the start and end code embedded in the TV signal, a number of TV stations all over the world do not support this feature. Therefore, existing DVRs are not able to get these kinds of programs fully recorded in this case. In this paper, we propose an "end detection" method based on video content analysis, which automatically detects whether the live program has reached its real end or not, thus the recording procedure will be kept before reaching the true end.

The proposed method is based on an observation that typically the video content within one particular sports program has relatively high consistency. To be exact, visually and/or aurally similar video segments will occur repeatedly through the entire program, while when a new program starts, this repetition breaks and may be replaced by another set of similar and repeating segments. The main reason of this phenomenon is that a sport game generally takes at one place (field), as well as the locations of the cameras for shooting the game are typically fixed. Based on above observation on content consistency, a two-round end detection scheme is proposed. The first round, *Candidate End Point Finding*, is able to find a coarse candidate end point, and the second round, *End Point Refining*, is able to refine the coarse result.

Closely related research work to the proposed scheme is similarity measure of video segments (may be scenes, shots, sub-shots, or any set of consecutive frames) in video retrieval area [1-3], as video segment comparison is the basis of examining video content consistency. In [3] low-level feature based signatures are applied to represent video segments, in which similarity of video segments depends on the distance between the corresponding signatures. In [1][2], segment similarity measure is built on shot level which takes the similarity between shots, temporal order, and granularity into account at the same time. In our work, the proposed scalable segment similarity measure is derived from typical shot-based similarity measure, while it is able to measure similarity in different levels, which has significantly improved the performance of the proposed end detection algorithm.

The rest of this paper is organized as follows. Section 2 describes an overview of TV end detection system. The scalable segment similarity measure is introduced in Section 3. Section 4 and Section 5 detail the two-round end detection scheme. Experiments and conclusion remarks are presented in Section 6 and 7, respectively.

2. SYSTEM OVERVIEW

The proposed end detection scheme consists of 3 stages: Shot Detection, Candidate End Point (EP) Finding and EP Refining. While the sports program is being broadcasted, DVR (or other devices with similar functionalities) starts to record the program as scheduled by users, and at the same time, a shot detector (similar to the one in [4]) is applied to detect the shot boundary. Thereafter, the

0-7803-9332-5/05/$20.00 ©2005 IEEE

recorded program could be represented as a shot sequence $S=\{s_1, s_2,\ldots, s_r, \ldots, s_m,\ldots, s_n,\ldots\}$, where s_r is the shot at the scheduled end of the sports program, while s_m is the true end shot, and s_n is the end shot of succeeding program after the current sports program (most likely the recording process will stop before reaching this shot, but it is presented here for convenience of system description). Accordingly the temporal structure of under-investigating TV program can be illustrated as Fig 1, in which R represents *reference program* (i.e., the part of the program within the scheduled time period), E stands for the part that the recording process need to be extended (i.e., the part between the *true end* and the *scheduled end*), while M and N denotes the whole sports program and the program after M, respectively. If we denote shot sequence $\{s_i, s_{i+1,\ldots,} s_j\}$ as $S(i, j)$, then we have

$$R=S(1, r),\ E=S(r+1, m),\ M=S(1, m),\ N=S(m+1, n) \quad (1)$$

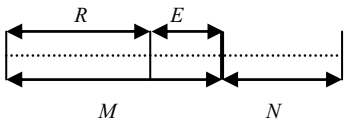

Fig 1. Temporal Structure of TV

The second stage of the system, Candidate EP Finding starts to work just after the broadcasting reaches the scheduled end. The recording procedure will not stop until candidate EP is found. Generally Candidate EP Finding needs to check about 15 more minutes after the real end point to ensure the real end point is not missed (as to be detailed in Section 4, a 15-minute sliding window is applied).

Finally, the third stage, EP Refining, is applied to find a more precise end point in the recorded program, and then the over-recorded part is cut out.

3. SCALABLE SEGMENT SIMILARITY

Although sports programs have relatively high content consistency, the degree of content consistency varies even within a single program, as well as between different types of sports programs. Therefore, simple threshold-based similarity rules are not adequate to describe the content consistency. For this reason, a scalable segment similarity measure is proposed to compare the TV segments in different similarity levels so that it is more immune to variations within and between programs.

The proposed similarity measure for comparing two TV segments (each segment is a series of shots) is derived from a typical shot similarity measure, in which a set of thresholds are applied to represent different similarity levels. For a certain similarity level z, only the shot pairs (one shot from each TV segment) whose similarity is higher than this similarity level contribute to the segment similarity. To be exact, if we define shot similarity of s_i and s_j as the normalized histogram intersection [1], denoted by $Sim(s_i, s_j)$, then the contribution of a shot pair to the segment similarity is defined as

$$Cont(s_i, s_j, z) = \begin{cases} Dur(s_i) \times Dur(s_j), & Sim(s_i, s_j) > z \\ 0, & Sim(s_i, s_j) \le z \end{cases} \quad (2)$$

where $Dur(s_i)$ is the duration of shot s_i. Finally, the segment similarity of segment $S(p_1, p_2)$ and $S(q_1, q_2)$ is defined as

$$Sim_{seg}(S(p_1, p_2), S(q_1, q_2), z) = \sum_{i=p_1}^{p_2} \sum_{j=q_1}^{q_2} Cont(s_i, s_j, z) \quad (3)$$

In our implementation, four different similarity levels are applied, *i.e., z* is set to 0.05, 0.10, 0.15 and 0.20, respectively. Practically before calculating segment similarity, dark (or nearly black) shots and short shots (less than 2 seconds) are filtered out due to all dark shots are similar in terms of color histogram, and short shots are often commercials. Because the comparison only uses color histograms, this similarity measure is not only fast but also memory-saving.

4. CANDIDATE END POINT FINDING

As aforementioned, content consistency of the recording TV program will break when the real end point of the sports video reaches. Based on this observation, Candidate EP Finding procedure tracks the change of segment similarity in a sliding window. The candidate end point will be found if the similarities fall under certain values.

As illustrated in Fig 2, Candidate EP Finding procedure starts after the scheduled end point with a lag period (i.e., a sliding window of 15 minutes, as to be explained). When a new shot is detected, the sliding window moves forward to cover the new shot, while the duration of the sliding window keeps unchanged.

Segment comparison between the reference program and the sliding window is taken on all similarity levels. If $Seg(k, l)$ denote the segment ending at shot s_k with a certain duration l, the sliding window can be denoted as $Seg(e, L_f)$, where e is the index of the newly detected shot and L_f is the length (duration) of sliding window. Then the segment similarity in the level z is calculated as

$$SS(z) = Max_{i<=r}(Sim_{seg}(Seg(i, L_R), Seg(e, L_f), z)) \quad (4)$$

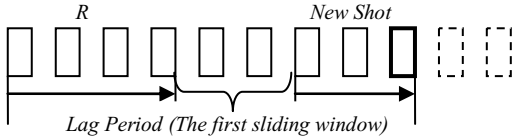

Fig 2. Candidate EP Finding

If the segment similarities in all levels are lower than predefined thresholds, the end shot of the corresponding sliding window is chosen as the candidate EP of the sports program. To be exact, the conditions are formalized as

$$SS(z) \le Low(z), \quad z = 0.05,0.10,0.15,0.20 \qquad (5)$$

In our experiments, the length of the sliding window is set to 15 minutes (L_f), and the duration of the reference segment is set 30 minutes (L_R). The values of *Low(z)* are set to the minimum values of *SS(z)* in a training data set which contains various types of sports programs.

5. END POINT REFINING

Due to a sliding window of 15 minutes is applied in candidate EP detection, the end point found in Section 4 generally will fall about 15 minutes behind the true end. In this step, a backward search algorithm is employed to find a more precise end point.

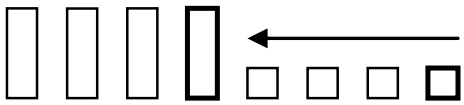

Fig 3. End Point Refining

We propose two EP Refining methods. The basic idea of the first method is similar to Candidate EP Finding, which is also based on segment comparison. It also uses a sliding window, but instead of moving the window forward to find consistency break, it moves the window backward from the candidate EP to find the first consistent segment. The length of sliding window is set to a relatively short value, say, 9 minutes, in order to achieve more precise result. The segment similarities are calculated similar to (4), while the stop conditions are

$$SS(z) > Low1(z) \qquad (6)$$

The second method is shot-based, which scans backward to find a representative shot in current broadcasted sports program. The content of a representative shot should reflect one of the main scenes in the sports program, and this kind of shot generally can be found in the last minute of the sports program. The advantage of this method is that if a representative shot appear near the true end, precise end point will be found.

To find representative shot in the end of the sports program, firstly the representative shots in the reference program have to be determined. This is accomplished by a dominant color based clustering algorithm that is applied on reference program. The shots closest to sufficiently large clusters are chosen as representative shots (the set of these shots are denoted by K). Then in the backward scanning process, the shot s satisfying below condition is regarded as a representative shot:

$$(\sum_{t \in K} cont(s,t,z)) > Low2(z) \qquad (7)$$

Similar to equation (5), the optimal values of z in (6) and (7) are obtained from a training set.

6. EXPERIMENTS

The proposed end detection algorithm is tested on a set of synthesized programs and real programs. In this section, firstly we define two performance evaluation measures for end point detection, and then show the evaluation results on synthesized programs and real programs respectively.

6.1 Evaluation of end detection

There are two criteria for a good end detection algorithm. First, the detected end should fall behind the true end so we can get the entire program recorded. Second, the detected end should be as close as possible to the real end. Accordingly two measures, *Pa* and *Pw(x)* are defined as follows:

$$Pa = |\{i: TE(i) \le DE(i),\ 1 \le i \le n\}| / n \qquad (8)$$

$$Pw(x) = |\{i: TE(i) \le DE(i) \le TE(i) + x,\ 1 \le i \le n\}| / n \qquad (9)$$

where n is the total number of test samples, $TE(i)$ is the moment of the true end and $DE(i)$ is the detected end time for the *i-th* test sample. Actually *Pa* counts the ratio of detected ends falling behind the true ends, while *Pw(x)* counts the ratio of detected ends falling behind the true ends with at most x minutes lag.

6.2 Testing on synthesized samples

Totally 2050 synthesized samples are generated by concatenating 41 sports programs with 50 miscellaneous programs, where the 41 sports programs include basketball, baseball, soccer, football, beach volleyball, racing, etc, and 50 miscellaneous programs include teleplay, film, cartoon, news, etc. The duration of synthesized samples vary from about 1 hour to 3 hours and the ground truth (true end point) of each synthesized video is the joint point of two programs. To evaluate the robustness of the end detection algorithm, only half of the program is applied as reference program. More precisely, each sub-segment in the sports program whose shot numbers is equal to half of that number of the entire program are taken as a test sample. That is, multiple end detections are applied on one synthesized video. For example, if a program has 1000 shots, then actually 501 sub-segments/samples (i.e., $S(i, i+499)$, $1 \le i \le 500$) are applied as reference programs to do end detection.

To obtain the optimal values for the parameters (z) in equation (5) (6) and (7), 20 randomly selected sports program from the above data set are concatenating with 50 miscellaneous programs to form a training data set, and the parameters are finely tuned on this set. This process is

repeated five times (thus the training set and test set will be changed each time). The evaluation results of Candidate EP Finding and EP Refining on training data set are shown in Table 1(a) and (b), respectively. Then we test the proposed algorithm on the rest of synthesized videos, while the results are shown in Table 2. It can be seen that average *Pw*(18) for candidate finding on training set is about 95% and the average *Pw*(18) on testing set is about 92%. Table 2 also shows that EP Refining algorithms have significantly improved the performance, especially when a relatively more accurate end point is required.

6.3 Testing On Real Programs

Real programs include 7 sports program and 1 Oscar award show (though award show is not sports video but it has similar properties as sports). The parameters used here is the same as the ones used in the above experiments. Similar to the method described in Section 6.2, to evaluate the robustness of the algorithm, sub-segments of the real program are applied as reference programs, thus the number of test samples is dramatically increased. From the evaluation results shown in Tab 3, we can see all the detected ends fall behind the true ends and most of them are within 9 minutes lag.

Table 1.(a) Performance of Candidate Finding on training set

#	*Pw*(9)	*Pw*(12)	*Pw*(15)	*Pw*(18)	*Pa*
1	0.214	0.366	0.886	**0.951**	**0.963**
2	0.257	0.477	0.908	**0.957**	**0.968**
3	0.191	0.393	0.833	**0.925**	**0.969**
4	0.262	0.504	0.915	**0.963**	**0.969**
5	0.300	0.502	0.922	**0.987**	**1.000**

Table 1.(b) Performance of EP Refining on training set

#	*Pw*(9)	*Pw*(12)	*Pw*(15)	*Pw*(18)	*Pa*
1	0.244	0.395	0.818	**0.953**	**1.000**
2	0.315	0.628	0.862	**0.901**	**0.901**
3	0.270	0.374	0.858	**0.979**	**1.000**
4	0.307	0.608	0.857	**0.902**	**0.908**
5	0.196	0.470	0.825	**0.883**	**0.888**

Table 2. Result on Synthesized samples

#	*Pw*(9)	*Pw*(12)	*Pw*(15)	*Pw*(18)	*Pa*
D1	**0.252**	0.487	0.877	**0.939**	**0.948**
D2	**0.909**	0.910	0.911	**0.914**	**0.943**
D3	**0.913**	0.925	0.929	**0.931**	**0.948**
D1 indicates Candidate EP Finding, D2 indicates segment-based EP Refining, and D3 indicates the shot-based EP Refining.					

Table 3.(a) Candidate EP Finding on real samples

#	Genre	Pw(9)	Pw(18)	Pa
1	NBA	0.000	1.000	1.000
2	NBA	0.000	1.000	1.000
3	NBA	0.000	0.000	1.000
4	NBA	0.000	1.000	1.000
5	Ping Pong	0.000	1.000	1.000
6	Soccer	0.000	1.000	1.000
7	Football	1.000	1.000	1.000
8	Oscar Award	0.000	1.000	1.000

Table 3.(b) EP Refining on real samples

		Segment-based		Shot-based	
#	Genre	*Pw*(9)	*Pa*	*Pw*(9)	*Pa*
1	NBA	1.000	1.000	1.000	1.000
2	NBA	1.000	1.000	0.016	1.000
3	NBA	0.000	1.000	0.000	1.000
4	NBA	0.677	1.000	0.888	1.000
5	Ping Pong	1.000	1.000	0.998	1.000
6	Soccer	1.000	1.000	1.000	1.000
7	Football	1.000	1.000	1.000	1.000
8	Oscar Award	0.998	1.000	1.000	1.000

The one with zero *Pw*(9) is due to an untypical repeating commercial shots with a relatively long duration, which occurs both frequently in the program and after the program. To solve this issue, a commercial detection algorithm [5] may be applied firstly before doing end detection.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an online automatic end detection scheme which enables DVRs to fully record live-broadcast sports programs when they run over time. The proposed scheme is based on the observation that typically the video content within one particular sports program has relatively high consistency. A scalable similarity measure and a two–round end detection scheme are proposed to find the break of this content consistency in the broadcasted TV program. Experiments show that the end detection algorithm is able to detect acceptable end points with at most 9 minutes lag while achieving high accuracy.

Future work may include exploring more properties for each kind of sports program so as to improve the accuracy of the algorithm, as well as testing on a larger set of real programs. We may also try to apply the scheme on other types of programs.

8. REFERENCES

[1] Yuxin Peng, Chong-Wah Ngo, "Clip-based similarity measure for hierarchical video retrieval", in *ACM Multimedia 2004*.

[2] A.K.Jain, A.Vailaya, W.Xiong, "Query by video clip", in A*CM Multimedia System*, *volume* 7, 1999.

[3] S.C.Cheung, A.Zakhor, "Efficient video similarity measurement with video signature", in *IEEE Transaction on Circuits and Systems for Video Technology*, 13(1), Jan 2003.

[4] H.-J. Zhang, A.Kankanhalli, S.W.Smoliar, "Automatic partitioning of full-motion video", *Multimedia Systems*, 1, 10-28, 1993.

[5] X.-S. Hua, L. Lu, H.-J. Zhang, "Robust Learning-Based TV Commercial Detection," *Intl Conf. on Multimedia and Expo 2005*.