AUTOMATICALLY DISCOVERING UNKNOWN SHORT VIDEO REPEATS

Xianfeng Yang[1], Ping Xue[1], Qi Tian[2]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University
[2]Institute for Infocomm Research, Singapore

ABSTRACT

In this paper we propose an efficient and robust method to automatically discover unknown short video repeats with arbitrary lengths, from a few seconds to a few minutes, from large video databases or streams. The proposed method consists of non-uniform video segmentation, self-similarity analysis, locality sensitive hashing, and video repeat boundary refinement. In order to achieve efficient and accurate processing feature extraction and similarity measure are performed at two levels: video frame level and video segment level. Experiments are conducted on 12 hour CNN/ABC news, and 12 hour documentaries (Discovery and National Geography)， high recall and precision of 98% - 99% have been achieved. Video repeats' boundaries can be located within several frames. Applying the proposed method for video structure analysis is also briefly discussed.

Index Terms: Multimedia computing, multimedia systems, video signal processing, database search, pattern recognition

1. INTRODUCTION

Repetitive short video clips contained in TV broadcasting streams often embed important structural and semantic information about the video content being broadcasted. For instance, important news segments shall be repeated with high frequency during a time. Commercials are another frequently repeated category that makes up of an important ingredient in TV program structure. Some other short video repeats interspersing in the stream may act as syntactic elements. For instance, program logos mark transition between different topics and reveal important clues for video structures. Sports replay flying logos generally symbolize forthcoming or outgoing important events. Their functionalities has driven us to exploit effective solution to identify short video repeats from video collections or streams, for the purpose of video structure analysis, important event mining, commercial detection and skipping etc.

A number of repeat video/audio clip identification methods have been proposed for specific applications such as TV commercial detection (usually above 10 - 30 seconds)[1,2,3], sports video replay logos (1 second or less)[4], news video content analysis and improved video compression [5], and FM radio station broadcasting content tracking[6]. Video repeat lengths vary, ranging from less than one second to a few minutes like TV commercials. Different video segmentation methods have been adapted to achieve for the applications; some are based on video shot [5]; some uses fixed size moving windows such as in Herley [6], it breaks audio streams into blocks with a minimum length (i.e. 30s); Cooper et al.[7] uses uniform sampling(every 10 frames) in their video self-similarity analysis. Shot based video segmentation is very useful to reduce the amounts of video segments to be matched to discover unknown video repeats, however, due to shot detection errors, especially, for video shots with gradual transitions, many short video repeats will be missed.

There are also extensive research works done in the past on searching *known video clips* from video database such as [2] to search given TV commercials, and [11] to search given clips from video database, which is so called content based video retrieval. To search unknown video repeats is more difficult than search known video repeats because we need to find video repeats at first. Once they are located, then the problem becomes to search known video repeats from large video databases or steams. In this paper we focus on automatically discovering unknown video repeats only.

The remainders of this paper are organized as follow: Section 2 - 4 presents our repeated clip identification approach. Section 5 briefly describes how the short video repeat discovery can be used in video structure analysis. Experimental results are presented in Section 6.

2. VIDEO REPEAT IDENTIFICATION FRAMEWORK

The proposed framework is shown in Fig.1. We employ two pass matching to identify repeated clips, with the first pass matching discovering potential repeated clips, and the second pass matching improving accuracy.

The first pass matching includes three temporal level video representations, namely video units (VU), video segments (VS) and video clips (VC), as well as corresponding video similarity measures. The first step is content based video segmentation. Video stream is

partitioned into basic video units (VU). The second step is self-similarity analysis. Video units are grouped by a window size *K*, i.e. two units as one group, to form bigger size video segments (VS), then they are compared with each other to produce similarity matrix **S.** By similarity measure f_1, two segments will be judged as either identical or non-identical, so **S** should a binary matrix. Here *locality-sensitive hashing (LSH)* is adopted to reduce correlation complexity. The third step is to identify repeat clips from similarity matrix S. Basically repeated clips can be identified from diagonals, which is controlled by similarity measure f_2.

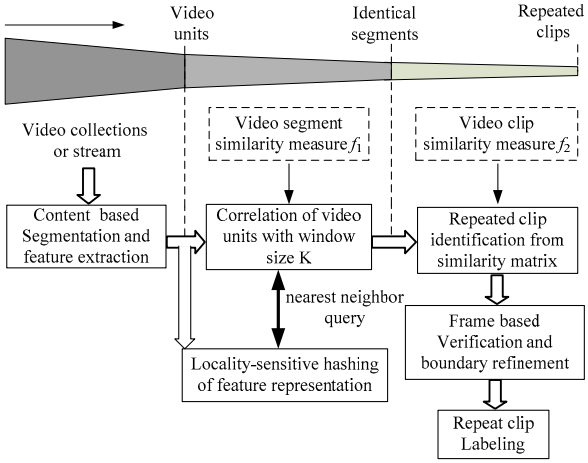


Fig 1: Framework for repeat video clip identification

The second pass matching adopts frame based matching to verify candidate repeated clips for accuracy improvement. After that a boundary refinement step is employed to extend repeated clips' boundaries close to their true ones as possible. The last step is repeated clip labeling. Repeated instances will be extracted from repeated clip pairs and grouped into multiple categories. Each category represents a unique repeat pattern.

3. VIDEO SEGMENTATION AND REPRESETNATION

In this Section we will discuss several existing video representation schemes, and propose ours.

3.1. Video representation overview

There are several ways to represent video for finding video repeats， including frame based, shot based, uniform sampling, and content based keyframe selection. *Continuous frame representation* is a faithful video representation without losing any temporal information, and there will be no sampling or segmentation errors. However, the large number of frames is difficult to manipulate and heavy computation load is needed. A simple video abstraction method is *uniform sampling* or segmentation, i.e. sampling every ten frames. However, since it is independent of video characteristics, content redundancy can not be well removed. Moreover, this technique can not adapt to boundary shift between to repeat objects.. *Shot based video representation* is also widely used in video analysis. Since it is consistent with video production process, it becomes a natural and popular way to exploit video structure. However for gradual transitions, which are widely used for many video repeats, accurate shot boundary detection remains problematic. Missing detection of these video shots means miss detection of the corresponding video repeats. *Content-based keyframe selection* is more efficient for video abstraction. It can well remove video redundancy while capturing significance of video sequence. Generally it can achieve more than ten times frame reduction, but at mean time it also leads to some temporal data loss

3.2. Keyframe based non-uniform segmentation and three level representation

In our method video stream is segmented by content-based keyframes, and interval between two consecutive keyframes is treated as the basic video unit (VU). Keyframe selection is based on color histogram difference. Suppose H_1 and H_0 are color histograms of current frame and the last keyframe respectively, then current frame is selected as new keyframe if

$$\left|1 - \text{inter}(H_1, H_0)\right| > \eta \qquad (1)$$

Where $\text{inter}(H_1, H_0)$ is intersection of two color histograms, η is threshold.

This representation is a seamless video segmentation without temporal data loss, which is similar to shot segmentation, but its granularity is smaller than shot. Its advantages lie in: First it is robust to boundary shift of repeat clips. Generally shift error can be corrected after a shot cut. Secondly it can reduce correlation between adjacent video units, so diagonal pattern will be sharper and easier be identified. The third advantage is that temporal length of video unit can be added to increase feature discrimination.

The second level video representation (VS) is formed by grouping two neighbor units. Compared to the first level, the second level has almost the same number of samples, but the discrimination ability will improve a lot, thus providing a less noisy output to build a higher level of video repeat clips.

3.3. Video features

Two types of video features are extracted. The first one is video unit (VU) feature used in first pass matching, and the other one is frame feature used in second pass matching.

Video unit (VU) feature is combination of color fingerprint proposed by us[8] and unit length. Color fingerprint is a fixed length string whose elements have six possible

symbols from {R, G, B, U, L, H}. It is robust to compression as well as moderate color distortions. Details about this color fingerprint please refer to [8]. In experiment we use one blending image for a unit and divide it into 8x8 blocks which result in 128-length string. We also apply LSH indexing on this color fingerprint, and its string representation can be easily transformed to a bit string required by LSH algorithm [9] without incurring extra errors. By LSH and unit length filtering, correlation complexity can be reduced by hundreds of times.

We also extract feature of each frame for second pass frame by frame matching. Each frame is divided into 4 sub-frames, and color histogram of each sub-frame is quantized to a symbol, so each frame is represented by 4 symbols.

4. VIDEO SIMILARITY MEASURES

Video similarity measures are conducted at several levels to ensure efficient and robust video repeat discovering: Video Unit, Video Segment, and Video Clip. Two video segments (VS) will be identical only if both of their two video units are identical in order, while similarity between VUs is measured by both color fingerprint and temporal length distance. Video similarity at Video Clip level is based on video frame features to refine video clip boundary in order to achieve high accuracy for boundary locations.

Finally, due to segmentation errors, repeated clips will appear as diagonals in similarity matrix, however, the diagonal line would be split into fragments. Moreover these fragments will not be collinear since non-uniform partition is used. Fig.2 shows a part of a similarity matrix computed in our experiment. As we can see diagonals are fragmented and contaminated by noises. To correctly get the whole repeated clip we design a hierarchical sequence linking and merging algorithm. The whole operation is totally based on temporal boundaries, so we need not process similarity matrix in memory whose pixel account is product of the numbers of video segments. Thus memory cost for this algorithm is low. More details can be found in [10].

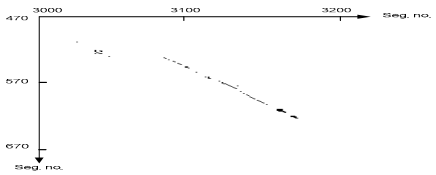


Fig 2: Example of diagonal pattern for repeated sequences

5. VIDEO STRUCTURE ANALYIS WITH SHORT VIDEO REPEAT IDENTIFICATION

Many TV stations use short video clips as structural video elements (SVE) before or after program sections (i.e. financial, sports news) in their regular daily or weekly broadcasting programs to indicate changes of topics, see Fig. 3. These program sections usually continue in the same manner for months or even years. If distribution of these short clips can be discovered, then corresponding video structure can be reconstructed. Since some short clips may repeat several times in one day's program, while others just repeat in different days, we need to collect programs in a fixed time period for several days to discover those repeat short clips.

Furthermore, TV channels uses SVE elements to indicate section boundaries within a program before and after TV commercial segments. By automatically detecting these SVE video programs can be segmented into video segments to construct table of content for video. We have proposed a model based on a directed graph to represent video structure in which video repeat detection is used to locate boundaries of video segments and then fit these video segments with the directed graph. Video structure can be reliably identified. Fig. 5 shows the news video structure analysis results for CNN news video, each vertical bar represents one half hour CNN news video, all of them have been segmented into individual sections such health, headline, dollar sense, travel, sports, etc. based on the SVEs. More results are presented in [10].

Fig 3. Examples of CNN program logos

6. EXPERIMENTAL RESULTS

For news video we chose half-hour CNN and ABC news videos from TRECVID data to form two video collections, each of which contains 12 day programs with 6 hours around. By manually searching short repeat clips including program logos and commercials, but neglecting other repeat scenes, i.e. anchor persons, 34 kinds of repeat clips with totally 186 instances are found from CNN collection, while 35 kinds with totally 116 instances found from ABC collection. For documentary video we used 12 hour video from Discovery Channel and National Discovery.

6.1. Recall and precision performance

After labeling we get 193 repeat instances from CNN collection. The shortest clip length is 0.8 seconds, while the longest one is 75 seconds. Recall of program logos and commercials is 184/186(98.9%), and precision is 98.4%. 124 repeat instances are extracted from ABC collections.

Recall of program logos and commercials is 100%, and precision is 96.8%. After frame by frame verification, precision can increase significantly, while recall just has slight drop. For instance, when recall and precision are 92% and 62% before verification, they become 91.6% and 99.6% respectively after verification. The equal error before verification is 14%, while that after verification can reach ~3%.

6.2. Boundary shift error

We selected 300 repeated clips that cover almost all labeled categories and checked their left and right boundary shift without boundary refinement. The smallest shift is 0 s, and the largest one is 16.4s. The average shift is 0.47s. We apply boundary refinement algorithm on some repeated pairs that have large shift errors, and found that the shift can be reduced to several frames.

6.3 Video structure discovery

Good performance in video structure identification has been achieved for both news video and documentary video. The Table 1 shows the SVE identification results for the 12 hour video content from Discovery and National Geography. Similar performance has been achieved for news video contents from CNN/ABC, because of space limitation our experimental results could not be included. Readers may find more details from our coming paper [10]

Table 1. Experimental results with documentaries.

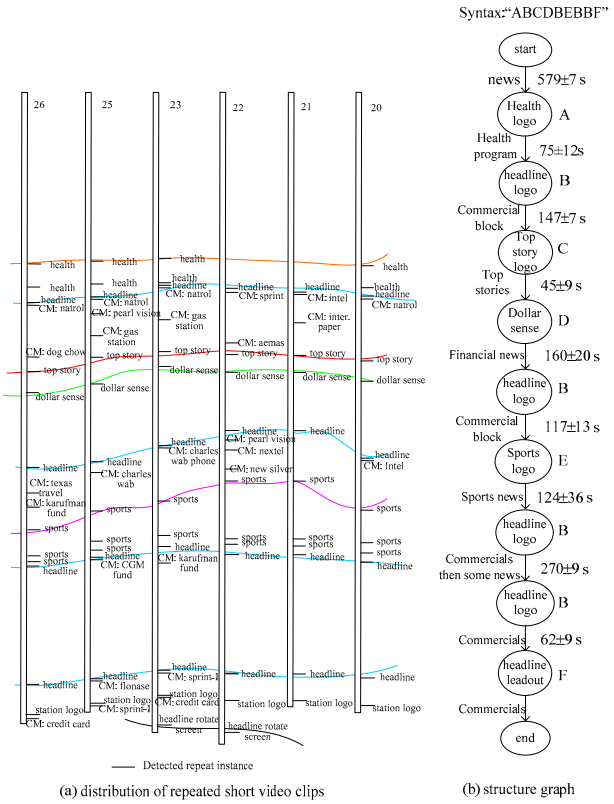
6.4. Detection speed on PCs

In a Pentium-4 2.5Ghz PC, the two-stage video repeats detection for 6 hour CNN takes only 22 seconds. Search of 1 min. long video from 12 hour video database can be completed in 1 second provided features are pre-calculated.

7. CONCLUSION

We have shown the proposed novel method can achieve efficient and robust unknown short video repeats discovering. The proposed segmentation method and the video similarity measures have achieved high precision and recall of 98-99%, on both news videos and documentary video.

Video Content ID	Video Length (in min.)	No. of SVE	Missed	False positive
Video II (NG)	47:00	4	0	0
Video V (NG)	1:13	10	2	0
Video I (DIS)	58:21	9	0	1
Video III (DIS)	58:26	9	0	0
Video IV (DIS)	1:00	9	0	0



(a) distribution of repeated short video clips (b) structure graph

Fig. 5. CNN news video structure identified for 6 half hour programs

8. REFERENCES

[1] L. Agnihotri, N. Dimitrova, T. McGee, S. Jeannin, D. Schaffer, J. Nesvadba, "Evolvable visual commercial detector", *Proc. IEEE CVPR*, 2003.

[2] R. Lienhart et al., "On the detection and Recognition of Television Commercials", *Proc. IEEE ICMCS,* 1997.

[3] S-C. Cheung, T P. Ngueyen, "Mining Arbitrary-length Repeated Patterns in Television Broadcast", *IEEE ICIP*, 2005.

[4] H. Pan, et al., "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions," *Proc. IEEE ICASSP,* 2002.

[5] K.M. Pua, J.M. Gauch et al. "Real time repeated video sequence identification", *CVIU, vol.* 93 (2004), pp.310-327.

[6] C. Herley, "Extracting Repeats from Media Streams", *Proc. IEEE ICASSP*, 2004.

[7] M. Cooper, J. Foote, "Scene Boundary Detection via Video Self-Similarity Analysis", IEEE ICIP, 2001.

[8]Xianfeng Yang, et al., "A Color Fingerprint of Video Shot for Content Identification", *Proc. ACM Multimedia-04*, pp. 276-279.

[9] A. Gionis, P. Indyky, R. Motwaniz, "Similarity Search in High Dimensions via Hashing", *Proc. Int'l. Conf. on Very Large Data Bases*, 1999, pp. 518—529.

[10] Xianfeng Yang, Qi Tian, Ping Xue, "Efficient Short Video Repeat Identification with Applications on News Video Structure Analysis", accepted by IEEE-T-MM, Sept. 2006

[11] A. K. Jain, et al., "Querying by Video Clip," ACM Multimedia Systems, vol. 7, pp. 369-384, 1999