SHOT-ADAPTIVE SAMPLING FOR HOME VIDEO SUMMARIZATION

Nagul COOHAROJANANANONE and Kiyoharu AIZAWA

Dept. of Electrical Engineering and Dept. of Frontier Informatics,
University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, JAPAN 113-8656
nagul@hal.t.u-tokyo.ac.jp

ABSTRACT

In this paper, we propose a new home video summarization by shot-based adaptive sampling using shot information. The characteristics of shots in video should be useful information for summarization algorithm. Our algorithm make use of shot information that are shot duration and shot motion activity. At each shot, a group of key-frames are sampled. The number of key-frames sampled from each shot is calculated from shot duration and shot motion activity. In the experiment, our algorithm and conventional key-frames sampling algorithms are compared. The density of the sampled key-frames in our algorithm is widely distributed. Thus, our method extracts more information than the existing methods.

1. INTRODUCTION

Recently, digital video cameras are widely used to record personal events such as vacations, weddings, and graduation ceremonies. These videos quickly accumulate to many hours of the video data. Nevertheless, playing home videos back or searching for one particular records tends to be too time-consuming. Therefore, it is more desirable to edit or summarize the entire records for guide scanning.

In this paper, we propose a video summarization that focuses to the home video data by adaptive sampling using shot information. Shot information which are shot duration and shot motion activity are applied to determine a number of key-frames representing each shot. Typically, a video composes of a collection of shots with different shot duration and shot motion. Summarization algorithm without consideration of shot boundary might result in insufficient information in summarized video. In this paper, we compare our algorithm to two non-shot based key-frames sampling algorithms. 1) Extracting key-frames by uniformly sampling the video frames from the original sequence at certain time intervals [1] [2]. 2) Adaptive sampling using R-sequence algorithm [3]. The R-sequence algorithm is very efficient summarization scheme, but it does not take into account shot information.

Summarization by representing key-frames is one of the most popular methods. There has been an interest in developing efficient schemes for video summarization based on key-frames extraction. Works in [4] [5] [6] [7] tried to extracted key-frames from clustered N groups without take into account shot boundary. Work in [8] divides the video sequence into n equal parts on the cumulative motion activity scale. Then, a frame at the middle of cumulative motion activity scale of each of the segments is set the key-frame. There are also interesting summarization works on using home video data. Algorithm proposed by Lienhart [9] creates a four-level cluster tree of the video clips based on the recorded date and time. The algorithm later randomly removes sub-trees at different levels to reach the desired output video length. The input by user is the total duration. Another algorithm proposed by [10] extracted a key-frame from a suitable shot using motion and brightness, however, a number of key frame does not depend on the shot duration therefore extracted key-frames may not hold enough information for the long shot. A use of shot duration and shot motion activity is reported in work based on tempo of the motion picture [11] Tempo is calculated by using shot duration and motion activity. Tempo's edge is later detected by the edge detector. However, the significant tempo changes often occur across the boundary of the story elements (scenes), therefore users are likely to receive information around the boundary of the story elements.

In the rest of the paper, the summarization algorithms are explained in Section 2. The experiment and the result are explained in Section 3. The conclusion will be described in Section 4.

2. SUMMARIZATION ALGORITHM

In this section, our method will be explained as well as the two existing method.

2.1. Uniformly sampling

Most of earlier work chose key frames by uniformly sampling the video frames from the original sequence at certain time intervals, which was applied in the Video Magnifier [1], MiniVideo system [2]. In order to sample a key-frame, A video sequence is equally divided into *N* blocks. In each blocks, the first, middle or the last frame is extracted as a key-frame. This is probably the simplest way to sample key frames without determining a shot boundary.

2.2. Adaptive sampling: R-sequence

An adaptive sampling algorithm was proposed by Sun and Kankanhalli [3] [12] where no shot detection is needed. On the contrary, the entire video sequence is first uniformly segmented into *L* units, and then a unit change value is computed for each unit, which equals to the frame difference between the first and last frame of the unit. Next, all the changes are sorted and classified into 2 clusters, the small-change cluster and the large change cluster, based on a predefined ratio r. Then for the units within the small-change cluster, the first and last frames are extracted as the R-frames, while for those in the large-change cluster, all frames are kept as the R-frames. Finally, if the desired number of key frames is obtained, the algorithm stops, otherwise, the retained R-frames are regrouped as a new video, and a new round of key frame selection is initialed. This method is considered as adaptive sampling taking into account the difference of the units. However, any shot detection is not included, and it tends to selected more key-frames only in moving scenes.

2.3. Shot-based adaptive sampling

Our proposed algorithm [13] make use of shot information that are shot duration and shot motion activity. Therefore, these two information need to be determined first. Shot duration can be determined by shot boundary. The shot boundary can be determined from the difference between the two consecutive frames using the city-block algorithm. Shot boundary is found between the frames that distance is larger than the threshold. Shot duration is a period between the first and the last frame in the shot. The shot, if the duration less than 100 frames, is omitted to remain only the meaningful shot. For shot motion activity, our algorithm determines motion activity from block matching algorithm or the motion information on the MPEG video.

Commonly, a longer shot holds more information than a shorter shot, which mean that longer shots should be represented by more number of key-frames in a summarized video. However, key-frames from longer shots with a low motion activity could have multiple frames with similar content. On the other hand, in the video sequence, a shot that its motion activity is greater than the average motion of the sequence is considered as containing highly dynamic scene, therefore the dynamic scene should be represented by more key-frames (more weight).

The number of key-frames of shot#i (N_{shot_i})is calculated by the following equation:

$$N_{shot_i} = W_{shot_i} \times TN_{shot_i}$$

where, W_{shot_i} is a weight parameter of shot#i and TN_{shot_i} is a temporary number of key-frames of shot#i.

$$W_{shot_i} = \begin{cases} \frac{(M_{shot_i} - M_{Avg})}{W_{max} - M_{Avg}} + 1, & M_{shot_i} > M_{Avg} \\ 0.5\frac{(M_{shot_i} - W_{min})}{M_{Avg} - W_{min}} + 0.5, & M_{shot_i} < M_{Avg} \end{cases}$$

where, M_{Avg} is the average motion activity of a sequence. M_{shot_i} is the average motion activity of each shot. The W_{max} and W_{min} are maximum weight and minimum weight respectively. They are set to 2 and 0.5 respectively. To calculate TN_{shot_i}, we first sort the entire shots in video sequence in descending order by their shot durations. The longest shot (rank = 1) has the largest number of key-frames. TN_{shot_i} is calculated by the following equation:

$$TN_{shot_i} = \frac{(Max - Min) \times (n_{shot} - i_{rank})}{(n_{shot} - 1)} + Min$$

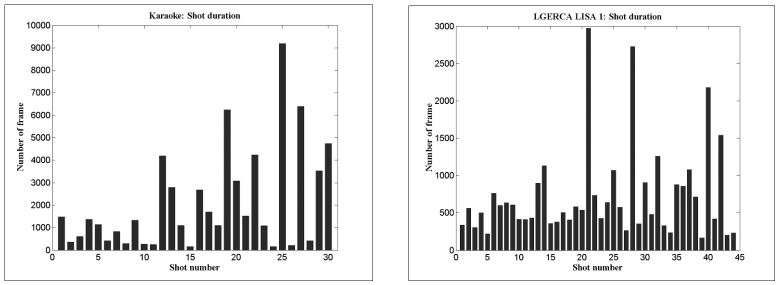
where, n_{shot} is the number of shots appearing in a summarized video. Max and Min are the maximum and minimum parameter for linear slope.

In next step, the key-frames are sampled from each shot according to desired number (TN_{shot_i}). Our algorithm sample high frame difference key-frames using the algorithm proposed in [3][12].

When the summarized video includes only the sampled frames, the result is not continuous. It is necessary to smooth the result by including the frames within S frames of R-frames. S frames is defined as smoothing parameter and given by users.

3. EXPERIMENT

In the experiment, three test home videos, (1) Karaoke (35 min) and (2) LGERCA LISA 1 (18 min) were summarized. Karaoke was about eating and sightseeing which included walking shots. Karaoke was recorded in the both outdoor and indoor. The event was approximately 5 hours long. LGERCA LISA 2 sequence is the video of a little girl in several events. Before sampling the key-frames, we determine the duration of shots from three sequences. During the shot boundary determination, the clips that contain hand shaking were grouped into small shots that their duration are less than 100 frames. These shots were not used in the calculation because they were omitted during the pre-processing. The shot duration after the pre-processing of



(a) (b)

Fig. 1. Graphs showing the duration of each in sequences (a) Karaoke (b) LGERSA LISA 1

three sequences are shown in figure 1. Our experiment further extracted the key-frames according three algorithms as we described in section 2. The key-frames are determined by the condition of 5% summary of each original sequence. The extracted key-frames are examined whether each key-frames belong which shot number. The examined results for three sequences are displayed in figure 2 and 4.

In figure 1, each of sequence contains a few long shots, where the rest of the shots are short. From the results in figure 2(a) and 4(a), most of key-frames were sampled from some few long shots. The rest of key-frames were extracted in small number from other shots. In figure 2(b) and 4(b), only some shots were represented by many key-frames and the rest of the shots were represented or were not represented by few key-frames. The results from our proposed algorithm in figure 2(c) and 4(c)) showed that the sampled key-frames were distributed in most of shots properly. From this, users are capable to receive more information of the original sequence from video summary.

4. CONCLUSION

This paper proposes a new sampling video algorithm for video summarization. The number of key-frames in each shot is controlled by the shot duration and motion activity. The experiments show that our algorithm gives better results than other two sampling algorithms.

5. REFERENCES

[1] M. Mills, "A magnifier tool for video data", *Proceeding ACM Human Computer Interface*, pp. 93-98, May. 1992.

[2] Y. Taniguchi, "An intuitive and efficient access interface to real-time incoming video based on automatic indexing", *Proceeding ACM Multimedia*, pp. 25-33, Nov. 1995

[3] X. Sun and M.S. Kankanhalli, "Video Summarization Using R-Sequences", *Journal of Real-Time Imaging*, vol. 6, pp. 449-459, Dec. 2000.

[4] Y. Zhuang, Y. Rui, T. S. Huang and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering", *Proceeding IEEE ICIP*, 1998.

[5] A. Hanjalic and H.J. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis", *Transaction on Circuits and Systems for Video Technology*, vol. 9, no. 8, Dec. 1999.

[6] A. Girgensohn and J. Boreczky, "Time-constrained keyframe selection technique", *Proceeding ICMS*, pp. 756-761, 1999.

[7] S. Uchihashi,J. Foote, A. Girgensohn, J. Boreczky" "Video manga: generating semantically meaningful video summaries", *Proceeding ACM Multimedia*, 1999.

[8] A. Divakaran, R. Radhakrishnan, K. Peker, "Motion Activity-Based Extraction of Key-Frames From Video Shots", *Proceeding ICIP*, Sept. 2002.

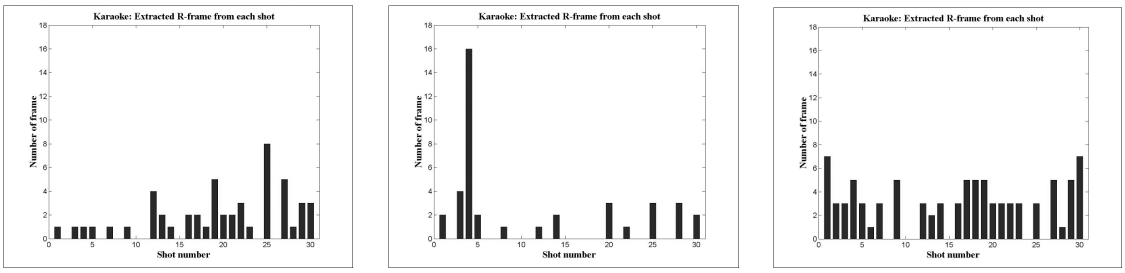
[9] R. Lienhart, "Dynamic Video Summarization of Home Video", *Proceeding IS&T/SPIE*, vol. 3972, pp. 378-389, 2000.

[10] A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, L. Wilcox, "A Semi-Automatic Approach to Home Video Editing", *Proceedings of UIST '00* pp. 81-89, 2000.

[11] B. Adams, C. Dora, S. Venkatesh, "Novel Approach Determining Tempo and Dramatic Story Sections in Motion Pictures", *Proceeding ICIP*, Sept. 2000.

[12] C. M. Chew and M. S. Kankanhalli, "Compressed domain summarization of digital video", *Proceeding IEEE Pacific-Rim Conference on Multimedia* pp. 490-497, 2001.

[13] N. Cooharojananone, S. Auethavekiat, K. Aizawa, "Home Video Summarization by Shot Characteristics and User's feedback", *To be appeared in Proceeding SPIE 2004* Jan 2004.



(a) (b) (c)

Fig. 2. Karaoke sequence: Graphs showing the sampled key-frames from each shot three algorithms (a) Uniform sampling (b) Adaptive sampling (R-sequence) (c) Shot-based adaptive sampling

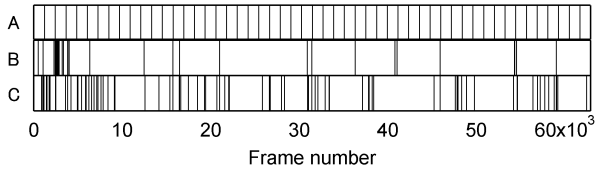
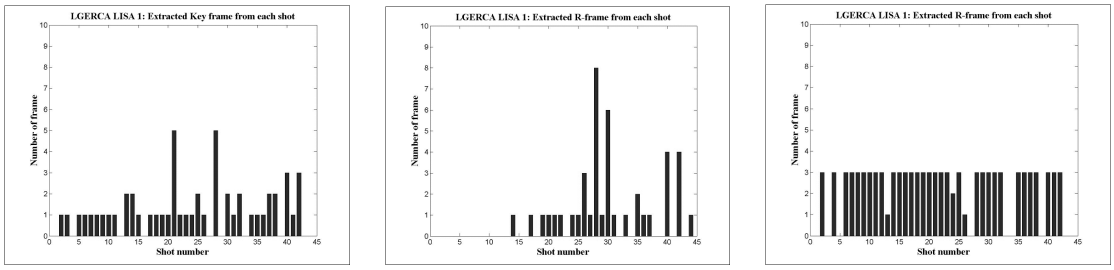


Fig. 3. Karaoke: Graphs showing the density of sampled key-frames from three algorithms (a) Uniform sampling (b) Adaptive sampling (R-sequence) (c) Shot-based adaptive sampling



(a) (b) (c)

Fig. 4. LGERCA LISA 1: Graphs showing the sampled key-frames from each shot three algorithms (a) Uniformly sampling (b) Adaptive sampling (R-sequence)(c) Shot-based adaptive sampling

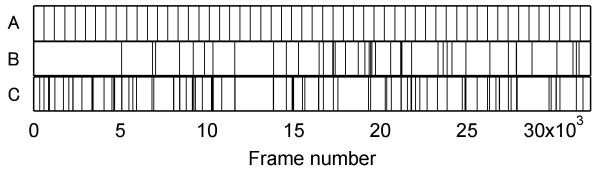


Fig. 5. LGERCA LISA 1: Graphs showing the density of sampled key-frames from three algorithms (a) Uniform sampling (b) Adaptive sampling (R-sequence) (c) Shot-based adaptive sampling