A RATE-CONSTRAINED KEY-FRAME EXTRACTION SCHEME FOR CHANNEL-AWARE VIDEO STREAMING

Yu-Hsuan Ho, *Wei-Ren Chen*, *and Chia-Wen Lin*

Department of Computer Science and Information Engineering,
National Chung Cheng University, Chiayi 621, Taiwan, R.O.C.
E-mail: cwlin@cs.ccu.edu.tw

ABSTRACT

This paper presents an adaptive rate-constrained key-frame selection scheme for channel-aware realtime video streaming applications. The proposed method dynamically determines the target number of key-frames by estimating the channel conditions according to feedback information. A two-step sequential key-frame selection scheme is then utilized to select a target number of key-frames by first finding the optimal allocation of key-frame budget among the video shots in a video clip using an analytical model, and then selecting most representative key-frames in each shot according to the allocation. The feature information used for key-frame selection is extracted offline and stored in the server as metadata for realtime streaming and transcoding. Experimental results show that the proposed method can achieve good performance with acceptable complexity.

1. INTRODUCTION

With the proliferation of online multimedia contents, the maturation of multimedia streaming systems, and the establishment of video coding standards, people can ubiquitously access and retrieve various multimedia contents through the Internet, promoting networked multimedia services at a fast pace. Since users are always spending time in repeatedly searching for a specific target object on the multi-media which has enormous video contents, it would be very helpful to develop convenient tools so that the users can browse and retrieve it effortlessly.

Video summarization [1-6] is a short summary of the content of a full-version video with a structured collection of selected frames, or key-frames, in such a way that the content is rapidly provided with concise information about the content while preserving the essential message of the original video. Instead of sending the whole video, a short-version video consisting of a relatively small number of key-frames can be used to meet the rate and time constraint for viewing the video in many practical applications. Besides visual summarization, key-frames also provide salient visual features (color, shape, and texture) for efficient video indexing and retrieval. In order to generate such a summary video, one major work is to identify which frames are the most representative key-frames. Another important issue is to determine an appropriate number of key-frames to well represent the whole video under the rate and viewing time constraints.

The simplest way of key-frame extraction is to use the first or the last frame of each shot as the shot's key-frame. Another simple method is to choose several key-frames separated by a fixed or random distance in a shot. Although sufficient for stationary shots, these methods are not adequate for dynamic shots since the visual contents may vary quickly. To obtain a good compact representation of video, most existing approaches tend to extract key-frames by adapting to the dynamical video content, which may be grouped into two categories [4]: cluster-based methods [1-2] and sequential-based methods [3-6].

The cluster-based methods take together and then cluster all the frames of a shot based on the similarity of their visual contents. As a result, one or more key frames, which are considered good representatives of the corresponding cluster, will be chosen from each cluster. For example, the method in [1] proposes to group the frames into several clusters according to their color histogram. In the case that the number of frames in a cluster exceeds a given number, this cluster will be represented by a set of key frames. Although such methods can provide a useful compact representation of the overall content of a video, they do not take into account the temporal information of frames, which is usually very important for understanding the video. On the other hand, the sequential-based methods [3-6] reduce the redundancy of temporal visual content by representing a number of consecutive frames in a shot by one key-frame extracted using both visual contents and temporal information of frames. For example, the method in [3] proposes to select the first frame of a video shot as a key frame, and then compare the following frames with the last key frame. If the dissimilarity between the two frames is large, this frame will be identified as a new key frame. A rate constrained scheme is proposed in [4] to reduce the temporal visual content redundancy by selecting a pre-determined number of key frames to minimize the representation distortion. This is achieved by iteratively making the positions of key frames and break-points locally optimal simultaneously. Similarly, two other cost-constrained methods based on the greedy and dynamic programming approaches are proposed in [5] and [6], respectively, for key-frame selection.

In this paper, we present an adaptive rate-constrained key-frame selection scheme for channel-aware realtime video streaming applications. The streaming server dynamically determines the target number of key-frames by estimating the channel conditions according to the feedback information. Under the constraint of the target key-frame number, a two-step sequential key-frame selection scheme is adopted to select the target number of key-frames by first finding the optimal allocation among the video shots in a video clip, and then selecting most representative key-frames in each shot according to the allocation. The feature information used for key-frame selection is extracted offline to meet the realtime constraint.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

The paper is organized as follows. Sec. 2 presents the proposed system architecture and describes the compress-domain feature extraction and distance metrics used in our work. Sec. 3 describes the proposed two-step key-frame selection scheme. The experimental results are shown in Sec. 4. Concluding remarks are drawn in Sec. 5.

2. CHANNEL-AWARE VIDEO STREAMING USING ADAPTIVE KEY-FRAME EXTRACTION

A. System Overview

In this work, we consider the application scenario of streaming of preencoded videos. Fig. 1 shows the server architecture of the proposed channel-aware video streaming system. The server offline extracts the compressed-domain video features (e.g., spatio-temporal distances between two neighboring frames) from the preencoded bit-stream, and then store the features in the server as auxiliary data (metadata) to guild the key-frame selection while performing real-time transcoding. A channel estimator based on the method described in [7] is implemented in the server to estimate the channel conditions according to the feedback channel statistics (e.g., packet loss rate and packet round-trip delay) and encoder buffer fullness. The extracted features and the estimated channel conditions are then used to determine the output bit-rate and the key-frame selection policy to guide the temporal-downscaling transcoding.

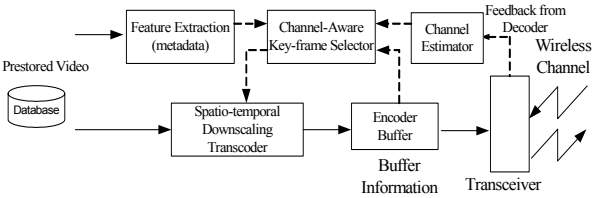


Fig. 1. Proposed streaming server architecture.

After determining the output bit-rate at some time instant, the target number of key-frames, N_{KF}, for representing a video clip with rate constraint R_{out} is obtained as:

$$N_{\mathrm{KF}} = \left\lfloor k_{\mathrm{adj}} \frac{N_{\mathrm{total}} R_{\mathrm{out}}}{R_{\mathrm{in}}} \right\rfloor \tag{1}$$

where R_{in} is the bit-rate of the incoming video; R_{out} is the estimated bitrate of the summary version of the output video; N_{total} is the length of the video clip; k_{adj} is a constant smaller than or equal to one to reflect the increase of coding complexity since the distance of two frames in the summary video becomes several times of that of the original video. In what follows, we will focus on the discussion of the proposed rate-constrained key-frame extraction scheme.

B. Compressed-Domain Feature Extraction and Distance Metrics

In order to identify what are the most representative key-frames in a video clip, we need to define distance metrics for measuring the distortions between a key-frame and those frames which are represented by the key-frame. Several compressed-domain features have been proposed for measuring the distance (dissimilarity) between two images [8]. In this work, we adopt the Hausdorff distance as a spatial dissimilarity measure since it has proven to be efficient [9] in characterizing the degree of mismatch between two edge images. Given two finite point sets $A = \{a_1,\ldots,a_p\}$ and $B = \{b_1,\ldots,b_q\}$, the Hausdorff distance is defined as [8]:

$$H(A,B) = \max(h(A,B), h(B,A)) \tag{2}$$

where

$$h(A,B) = \max_{a\in A} \min_{b\in B} \|a-b\| \tag{3}$$

and $\|\bullet\|$ is some underlying norm metric on the points of A and B (e.g., the L_2 or Euclidean norm).

Because the Hausdorff distance is sensitive to noise, a modified Hausdorff distance metric was proposed in [9] to compare the similarity of one edge image with part of edge images to mitigate the noise effect as follows.

$$h_K(A,B) = K^{\mathrm{th}}_{a\in A} \min_{b\in B} \|a-b\| \tag{4}$$

where K^{th} denotes the Kth ranked value in the set of distance (one corresponding to each element of B).

In this work, we extract the DC image of each frame in the compressed domain, then apply the Canny edge operator to extract the Canny edgy images of the DCT images. The spatial distance between the ith and jth edge images are defined as the cumulative modified Hausdorff distance as follows:

$$d^{\mathrm{S}}(f_i, f_j) = \sum_{n=i}^{j-1} h_K(f_n, f_{n+1}) \tag{5}$$

To take into account the object and camera motions, we define the temporal distance as the cumulative sum of motion vector magnitudes.

$$d^{\mathrm{T}}(f_i, f_j) = \sum_{n=i+1}^{j} \sum_{m=1}^{N_{\mathrm{MB}}} \left|MVX_m^n\right| + \left|MVY_m^n\right| \tag{6}$$

where MVX_m^n and MVY_m^n are the horizontal and vertical components of the motion vector of the mth macroblock of the nth frame; N_{MB} is the number of macroblocks in a frame.

We then combine the above two distance metrics to obtain a spatio-temporal distance as follows:

$$d(f_i, f_j) = k_{\mathrm{T}} d^{\mathrm{T}}(f_i, f_j) + k_{\mathrm{S}} d^{\mathrm{S}}(f_i, f_j) \tag{7}$$

where k_{T} and k_{S} are two weights which are determined empirically.

3. PROPOSED TWO-STEP RATE-CONSTRAINED KEY-FRAME EXTRACTION SCHEME

A. R-D Optimized Shot-Level Key-Frame Allocation

After computing the target key-frame number (N_{KF}) of a video clip, the problem of key-frame allocation is to distribute N_{KF} key-frames into all shots in the video clip. This is basically a resource allocation problem. Since different shots may have different characteristics (e.g., motion and texture), simply uniformly distributing the key-frames into the shots of a video clip will result in non-uniform representation quality. In the following, we consider the problem of how to achieve an optimal shot-level key-frame allocation so that the overall representation distortion of the video clip can be minimized. To achieve this goal, we propose an R-D optimized shot-level key-frame allocation scheme.

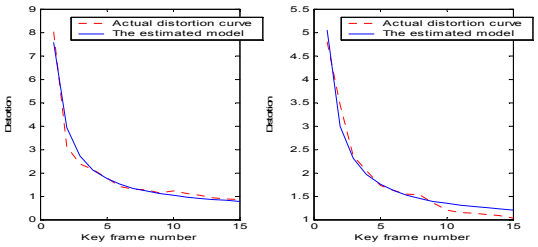
Assume that a video clip has N_{shot} shots, and the ith shot has its average representation distortion function $D_i(N_i)$ to characterize

the relationship between the representation distortion (D_i) of a video shot and the number of key-frames (N_i) used to represent the shot. We found the following first-order model is a good approximation of $D_i(N_i)$:

$$D_i(N_i) = \frac{a_i}{N_i} + b_i \tag{8}$$

where D_i stands for the average distortion of the ith shot when N_i key-frames are used to represent the shot; a_i and b_i are the model parameters estimated from a number of measured $D_i(N_i)$ values by using the least-squares estimation.

Fig. 2 compares the actual distortion curves and their model estimates obtained by (8) for two shots. This simple model has promising accuracy and is mathematically tractable when being used for finding the optimal key-frame allocation, as will be elaborated later.



(a) video shot 1 (b) video shot 2

Fig. 2. The actual distortion curves and modeled R-D curves.

With the model in (8), the optimal key-frame allocation problem can be formulated as the following constrained optimization problem:

$$\min \sum_{i=1}^{N_{\text{shot}}} L_i D_i(N_i) \quad \text{subject to} \quad \sum_{i=1}^{N_{\text{shot}}} N_i = N_{\text{KF}} \tag{9}$$

where L_i is the length of the ith shot.

The Lagrange multiplier can then be used to convert Eq. (9) to the following unconstrained optimization problem.

$$\min f = \sum_{i=1}^{N_{\text{shot}}} L_i\left(\frac{a_i}{N_i} + b_i\right) + \lambda\left(\sum_{i=1}^{N_{\text{shot}}} N_i - N_{\text{KF}}\right) \tag{10}$$

where λ is the Lagrange multiplier.

By setting partial derivatives to zero (i.e., $\partial f / \partial N_i = 0$ and $\partial f / \partial \lambda = 0$), the solution to Eq. (10) is obtained as follows:

$$\lambda = \frac{\left(\sum_{i=1}^{N_{\text{shot}}} \sqrt{L_i a_i}\right)^2}{N_{\text{KF}}^2} \quad \text{and} \quad N_i = \sqrt{\frac{L_i a_i}{\lambda}} \tag{11}$$

B. Key-frame selection within a video shot

After obtaining the key-frame allocation of each shot, the next step is to extract the most representative key-frames among the frames in each video shot so as to minimize the representation distortion. Fig. 3 illustrates an example of extracting N key-frames from a video shot of M frames, where $N = 4$ and $M = 22$ in this example. Let $\mathbf{F} = \{f_1, f_2, \ldots, f_M\}$ be the set of the frames in the video shot, $\mathbf{K} = \{k_1, k_2, \ldots, k_N\}$ the set of extracted key-frames. Then we can partition the shot $\mathbf{F}$ into a set of N non-overlapping intervals $\mathbf{T} = \{T_1, T_2, \ldots, T_N\} = \{t_0 \sim t_1, t_1 \sim t_2, \ldots, t_{N-1} \sim t_N\}$ such that the frames in the ith interval, $T_i = t_{i-1} \sim t_i$, are all represented by the ith key-frame, k_i. The right-boundary frame of each interval is called a break point. We then define the distortion of the ith interval represented by the ith key-frame as the sum of distance values between the key-frame k_i and the remaining frames in T_i as follows.

$$D(T_i, k_i) = \sum_{t=t_{i-1}}^{t_i} d(t, k_i) \tag{12}$$

where $d(t,k_i)$ is the distance value between frames t and k_i calculated by Eq. (7). The goal of this step is to find an optimal combination of $\mathbf{T}$ and $\mathbf{K}$ so that the overall distortion, $\sum_i D(T_i, k_i)$, can be minimized. However, this usually does not have a close-form solution, since it is difficult to find a mathematically tractable model to characterize the distortion function with a reasonable accuracy. In [4], an iterative procedure is proposed to find the optimal key-frame extraction from a video shot. In [6], a dynamic programming based scheme is proposed. However, both schemes may be computationally too expensive to be applied in real-time applications, especially when the number of clients becomes large. Another greedy algorithm is proposed in [5] for key-frame selection, which has lower complexity, but leads to higher distortion.

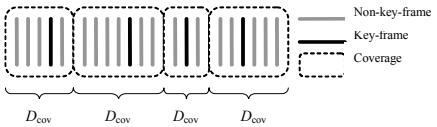


Fig. 3. Illustration of key-frame selection in a video shot.

In this work, we propose a low-complexity approach as follows:

Step 1. Summing up the distance values of all frames in a shot and dividing it by N, we can get the average coverage range of each key-frame as defined below.

$$D_{\text{cov}} = \frac{1}{N} \sum_{n=1}^{M-1} d(f_n, f_{n+1}) \tag{13}$$

The video shot is then partitioned into N intervals $\{t_0 \sim t_1, t_1 \sim t_2, \ldots, t_{N-1} \sim t_N\}$ each with approximately equal average representation distortion of D_{cov} as follows:

$$\sum_{n=t_{j-1}}^{t_j - 1} d(f_n, f_{n+1}) \cong D_{\text{cov}} \quad j = 1, 2, \ldots, N \tag{14}$$

Step 2. For each interval T_i, the key-frame is selected so as to minimize the distortion $D(T_i,k_i)$:

$$k_i = \arg \min_{t_{i-1} \le k_i \le t_i} D(T_i, k_i), \quad i = 1, 2, \ldots, N \tag{15}$$

4. EXPERIMENTAL RESULTS

Two CIF movie video clips ("Terminator III" and "The Lion Roars") each consists of four shots with clip-lengths of 727 and 696 frames respectively are used in our experiments. The two video clips are encoded at 30 fps and 1Mbps using an MPEG-4 encoder. Fig. 4 compares the average representation distortion of selecting key-frames from two selected video shots using the uniform selection scheme and the proposed scheme mentioned in Sec. 3.B. By minimizing the local distortion $D(T_i,k_i)$ based on equal-distortion interval selection (break-points), the proposed

scheme achieves lower distortion than the uniform key-frame extraction.

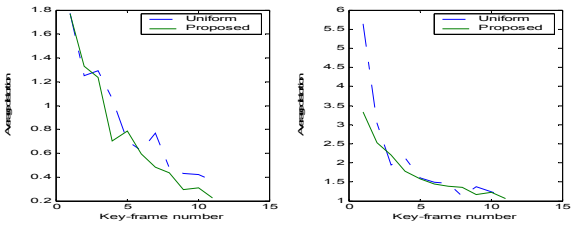


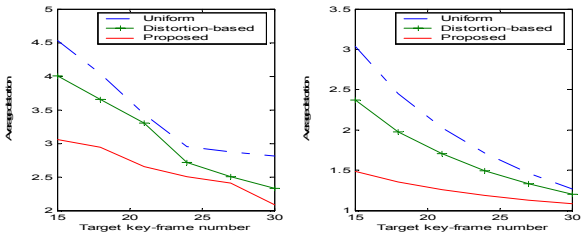
Fig. 4. Curve of distortion between frames and selected key-frames.

In addition to the proposed rate-constrained key-frame allocation scheme in Eq. (11), we also implement two other allocation schemes: the uniform allocation in Eq. (16) and the distortion-weighted allocation in Eq. (17), respectively, for performance comparison.

$$N_i = N_{\mathrm{KF}} \times \frac{L_i}{\sum_{i=1}^{N_{\mathrm{shot}}} L_i} \tag{16}$$

$$N_i = N_{\mathrm{KF}} \times \frac{L_i D_i}{\sum_{i=1}^{n} L_i D_i} \tag{17}$$

The uniform allocation scheme selects key-frames with uniform temporal distances, while the distortion-weighted approach selects key-frames with uniform distortion coverage. Fig. 5 shows the performance comparison of the uniform, distortion-weighted, and proposed key-frame selection methods in terms of average representation distortion for the two test video shots, respectively. The number of key-frames allocated for each shot ranges from 15 to 30 frames. Because our rate-constrained scheme is aimed at finding the optimal allocation of key-frames for different shots according to the characteristics of each shot, the resultant average distortion is significantly lower than those of the other two methods.



(a) video clip #1 (b) video clip #2

Fig. 5. Comparison of the average distortion values between uniform allocation, distortion-weighted allocation and R-D optimized allocation methods.

Fig. 6 shows the key-frames extracted from form the second shot of the "Terminator III" clip using the three methods. Six key-frames are selected from this 559-frame shot, resulting in the average representation distortion values of 3.85, 3.64, and 3.20 for the uniform allocation, the distortion-weighted allocation, and the proposed allocation methods, respectively. The proposed method can extract a better compact representation of the original clip.

Fig. 6. Key-frames (N = 6) selected form one shot of the "Terminator III" clip using the uniform (top-row), distortion-weighted (middle-row), and the proposed methods (bottom-row).

5. CONCLUSION

In this paper, we proposed an adaptive sequential key-frame selection method for channel-aware video streaming applications. The proposed two-step key-frame selection method first finds the optimal allocation among the video shots in a video clip by using a first-order analytical model, and then selects the key-frames in each shot according to the allocation. The Hausdorff distance values of two neighboring DC edge images, and the motion vector magnitudes are offline extracted and stored in the server as the feature information for realtime key-frame selection in streaming. Experimental results demonstrate the effectiveness of the proposed method.

REFERENCES

[1] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, pp. 866–870, Oct. 1998, Chicago, IL.

[2] A. Hanjalic and H. Zhang, "An integrated scheme for automatic video abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1280-1289, Dec. 1999.

[3] M. Yeung and B. Liu, "Efficient matching and clustering of video shots," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, pp. 338-341, Dec. 1995. Washington D.C.

[4] H.-C. Lee and S.-D. Kim, "Iterative key frame selection in the rate-constraint environment," *Signal Processing*: *Image Commun.*, vol. 18, pp. 1-15, 2003.

[5] Z. Li, A. Katsaggelos, and B. Gandhi, "Temporal rate-distortion optimal video summary generation," in *Proc. IEEE Conf. Multimedia & Expo*, pp. 693-696, July 2003, Baltimore, MD.

[6] X. S. Zhou and S.-P. Liou, "Optimal nonlinear sampling for video streaming at low bit rates," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 535-545, June 2002.

[7] S. Aramvith, I.-M. Pao, and M.-T. Sun, "A rate control scheme for video transport over wireless channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no.5, pp.569-580, May. 2001.

[8] H. Wang et al., "Survey of compressed-domain features used in audio-visual indexing and analysis," *J. Vis. Commun. Image R.*, vol. 14, no. 2, pp.150-183, June 2003.

[9] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 9. Sep. 1993.