

Iterative key frame selection in the rate-constraint environment

Hun-Cheol Lee^{a,*}, Seong-Dae Kim^b

^a SAT Group, Digital TV Laboratory, LG Electronics, San 4-1, Bongcheon7-Dong, Kwanak-Gu, Seoul 151-818, South Korea

^b Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, 373-1, Kusong-Dong, Yusong-Gu, Daejeon 305-701, South Korea

Received 19 February 2002; received in revised form 17 July 2002; accepted 10 August 2002

Abstract

Video representation through key frames has been addressed frequently as an efficient way of preserving the whole temporal information of sequence with a considerably smaller amount of data. Such compact video representation is suitable for the purpose of video browsing in limited storage or transmission bandwidth environments. In this case, the controllability of the total key frame number (i.e. key frame rate) depending on the storage or bandwidth capacity is an important requirement for the key frame selection method. In this paper, we present a sequential key frame selection method when the number of key frames is given as a constraint. It first selects the pre-determined number of initial key frames and time-intervals. Then, it adjusts the positions of key frames and time-intervals by iteration, which reduces the distortion step by step. Experimental results demonstrate the improved performance of our algorithm over the existing approaches.

© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Video indexing; Key frame selection; Rate-constrained; Video retrieval; Visual content representation

1. Introduction

Recently, many researches and standards have been proposed for the effective representation and retrieval of digital video [6,10,18]. Such trends are accelerated due to the emergence of interactive digital TV, the production of multimedia material distributed on CD-ROMs and the introduction of the new DVD technology. Key frames are a reduced set of still images that can represent the visual content of a video. The usefulness of the key frames lies not only in the compact representation of video sequence [16,17]. The key frames can also

be described by their visual features and used in video retrieval applications [13,14] as well.

The first step commonly taken for content-based digital video representation and retrieval is to detect shot boundaries [1,4,6,9]. After the shot boundaries are identified, key frames are selected in each shot. With these key frames, most of the existing works compute the similarity between two video segments based on the key frames in order to retrieve desired video segments [4,14] or represent a video in a scene-transitive way using such tools as video poster or scene-transition graph [16,17]. In this way, video representation and retrieval through key frames has been addressed frequently as an efficient way of preserving the whole visual and temporal information of sequence with

*Corresponding author.

E-mail address: lhc@sdvision.kaist.ac.kr (H.-C. Lee).

a considerably smaller amount of data. Such compact video representation is suitable especially when the available resources are limited. For example, in a mobile VOD service system, various summary sequences can be made with different numbers of key frames depending on the transmission bandwidth or storage capacity of a mobile terminal, which aim to reduce a long video into a short sequence in order to help the user to determine if a video program is worth viewing in its entirety. Like this, the controllability of the total key frame number (i.e. key frame rate) depending on the resources (storage or bandwidth capacity) is an important requirement for the key frame selection method. In addition, such functionality is useful in constructing a hierarchical video browsing system, where, depending on the hierarchy, different numbers of key frames are needed.

In this paper, we present a sequential key frame selection method when the number of key frames is given as a constraint. The key frame selection is modeled as collecting a set of representative frames in the sense of reducing the distortion in Fig. 1 as much as possible, while keeping the number of key frames constant.

This paper is organized as follows. In Section 2, we review some of the representative previous approaches and their drawbacks, and then we present our proposed method in Section 3. First, the distortion is defined and iterative key frame extraction method is explained, followed by the proof of its convergence and the procedure of

selecting the initial key frames and initial time-intervals. Section 4 shows the experimental evaluation, and Section 5 concludes this paper.

2. Review of key frame selection methods

The simplest way of key frame extraction is to use the first or last frame of each shot as the shot's key frame [11]. Another simple method is to choose several key frames separated by a fixed distance in a shot [12]. Although sufficient for stationary shots, these methods are not adequate for dynamic shots since the visual content of a shot is not considered at all. To overcome these defects, most existing approaches are based on visual content of video sequence. They may be grouped into two categories depending on the use of temporal information: cluster-based methods [2,3,19] and sequential methods [5,7,8,15].

In the cluster-based method, all the frames of a shot are taken together and clustered based on the similarity of their visual content. The key frame set is then obtained by collecting representatives of each cluster. In this method, temporal information of frames is not considered at all, namely, key frames are selected regardless of the temporal order of each frame. Zhuang et al. [19] represent each frame as a point in a multi-dimensional feature space using color histogram and group the frames into several clusters. A key frame set is finally obtained by collecting the representatives of the clusters that contain more than a given number of frames. Doulamis et al. [3] formulate feature vectors of frames based on color and motion segmentation. Then, the key frame selection is performed using an optimization method for locating a set of minimally correlated feature vectors. Chang et al. [2] consider the key frame extraction problem from a set-theoretic point of view, and derive systematic algorithms to find a minimal set of key frames that satisfies the given distortion bound.

Although the key frames extracted by the cluster-based methods are useful to understand the overall visual content of video, these key frames are not suitable for computing the similarity between two video segments since the temporal

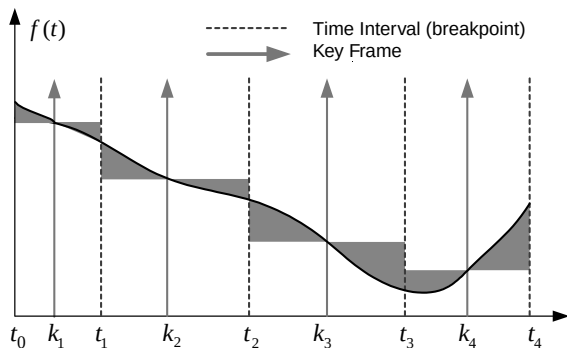


Fig. 1. Illustration of frame feature vector $f(t)$, key frames k_i , breakpoints t_i and distortion (in case of $N = 4$).

information of the video frames is not taken into account, which is an important property of a *video*.

On the other hand, the sequential key frame selection methods use both visual and temporal information of frames based on the criterion of reducing the temporal visual-content redundancy by representing several consecutive frames by one key frame. Yeung and Liu [15] initially select the first frame of a shot as a key frame and then compare the following frames with the last key frame. If the dissimilarity is larger than a threshold, then that frame is selected as a new key frame. Although simple, it is heuristic and the resulting key frame number is dependent on the threshold value.

Hanjalic et al. [5,7] propose a more sophisticated key frame allocation method, which aims to extract the pre-determined number of key frames. For example, when N is assigned as the number of key frames to be selected in a shot, they try to minimize the L_1 -distortion, which is a function of key frames and breakpoints, as shown in Fig. 1.

Here, they use accumulated frame-by-frame difference as a frame descriptor $f(t)$ for the image frame displayed at time t . By denoting the i th key frame as k_i and the breakpoint between k_i and k_{i+1} as t_i , where t_0 is the temporal starting point of a shot, they try to make every key frame locally optimal between two breakpoints and, simultaneously, every breakpoint locally optimal between two consecutive key frames in two steps: First, after assuming that all breakpoints are given, the key frame k_i between t_{i-1} and t_i is computed as (1) so that the first derivative of distortion function with respect to k_i may become zero.

$$k_i = \frac{1}{2}(t_{i-1} + t_i), \quad i = 1, 2, \dots, N. \quad (1)$$

Then, the position of a breakpoint between two consecutive key frames is obtained as (2), by which the first derivative of distortion function with respect to each breakpoint becomes zero.

$$f(t_i) = \frac{1}{2}(f(k_i) + f(k_{i+1})), \quad i = 1, 2, \dots, N-1. \quad (2)$$

From (1) and (2), $2N-1$ sets of equations are obtained for $2N-1$ unknowns ($t_1, t_2, \dots, t_{N-1}, k_1, k_2, \dots, k_N$). To solve these equations, Hanjalic et al. propose a recursive search algorithm using (1) and (2). They arbitrarily initialize the first

breakpoint (t_1) and compute the positions of following key frames and breakpoints in a recursive way. Starting with the assumed breakpoint t_1 , key frame k_1 is computed using (1). Then, from (2), k_2 is computed. Subsequently, from k_2 , t_2 is obtained using (1) and so on. In this method, the last breakpoint (t_N) should be identical to the given length of a shot. The authors emphasize that by adjusting the value of t_1 depending on the mismatch between the computed t_N and the actual length of a shot, the pre-determined number of key frames can be extracted.

They also claim that their recursive search method can obtain an optimal solution since, in deriving (1) and (2), the first derivative of the distortion function with respect to each unknown becomes zero. Of course, it is evident that the solutions (1) and (2) are locally optimal. The zero-first derivative, however, is a necessary condition for global minimization, namely, the solutions that are found by the recursive method may be local minimum, local maximum or even global maximum in the worst case.

Also, the recursive search method of Hanjalic's method inevitably necessitates the specification of first breakpoint. Therefore, like Yeung and Liu's method, many trial-and-errors are needed in order to extract the predefined number of key frames.

To avoid such trial-and-errors, in our earlier work [8], we proposed a rate-constrained key frame selection method, which does not need any threshold value or initial parameter specification. In this method, we select the pre-determined number of key frames independently with one another under the assumption that the previous key frame does not influence the position of current key frame, while making the total magnitude of temporal variations within a video shot be equally divided among the key frames. Although this method selects locally optimal key frames, the resulting breakpoints are not locally optimal.

3. Proposed method

To improve the method presented in [8], we propose an iteration-based key frame selection

method which selects the pre-determined number of key frames, while reducing the temporal visual content redundancy as much as possible and making the positions of key frames and breakpoints locally optimal simultaneously. In this section, we first define a distortion and formulate our problem quantitatively. Then, we explain the iteration-based key frame selection method, followed by the proof of its convergence and the selection of initial key frames.

3.1. Distortion measures of selected key frames

For the measurement of the temporal variation and distortion, appropriate visual features should be selected in order to represent each frame of a video sequence by a one (or multi)-dimensional feature vector. The visual feature can be composed using texture, color, intensity, motion, shape information, or any combination of these. Generally, visual feature itself is used as elements of a feature vector without any modification [2,3,8,15,19], but sometimes, cumulative absolute frame-by-frame feature difference is used as an element of a frame feature vector [5,7]. It is noteworthy that, depending on the user-interest, different visual features are chosen. For example, if the user is interested in the motion and wishes to select key frames more densely in the region of large activity, he (she) may select the average magnitude of motion vectors as a visual feature. If the user is interested in the intensity, the average or variance of pixel intensity may be chosen as a visual feature.

Let $\mathbf{S} = \{s_1, s_2, \dots, s_M\}$ be a set of M frames in a video shot and $\mathbf{K} = \{k_1, k_2, \dots, k_N\}$ be a set of N key frames of this shot. Here N , which is given in advance, is the pre-determined number of key frames, s_i is the i th frame in the shot and k_j is the j th key frame. Note that $\mathbf{K} \subset \mathbf{S}$ and $N < M$. For the simplicity of mathematical notation and derivation, we will consider each element of $\mathbf{S}$ and $\mathbf{K}$ as its temporal position, which is a continuous variable, although discrete version will be used in the practical implementation. Let $f(t)$ be a frame feature vector for the image frame displayed at time t and let $\mathbf{T} = \{[t_{i-1}, t_i] | i = 1, 2, \dots, N\}$ be a set of N time-intervals in each of which only one key

frame can exist. This means that the i th key frame k_i represents all the frames displayed during the time-interval $[t_{i-1}, t_i]$, which inevitably gives rise to a distortion as illustrated in Fig. 1. Let $d(t_i, t_j)$ be a dissimilarity measure between two frames located at time t_i and t_j , the form of which is application-dependent. Then, the distortion $D(\mathbf{K}, \mathbf{T})$ can be formulated in two different ways as follows. Note that these two formulas are same in the meaning, but have different forms just for the ease of mathematical derivation and explanation.

In the first formula, we consider the distortion $D(\mathbf{K}, \mathbf{T})$ as the sum of local distortions occurring in the non-overlapping time-intervals as (3).

$$D(\mathbf{K}, \mathbf{T}) \equiv D_1(\mathbf{K}, \mathbf{T}) = \sum_{i=1}^N \int_{t_{i-1}}^{t_i} d(t, k_i) dt. \quad (3)$$

Here, t_0 and t_N are the temporal positions of the first and last frame of a shot, respectively. This formula can be used to find a set of key frames that minimizes the distortion when the time-intervals are given in advance. A key frame set that minimizes (3) is obtained in each time-interval as (4). Of course, depending on $d(\cdot)$ and frame feature vector, a more analytic solution like (1) can be obtained.

$$k_i = \arg \min_{t_{i-1} \leq k_i \leq t_i} \int_{t_{i-1}}^{t_i} d(t, k_i) dt, \quad i = 1, 2, \dots, N. \quad (4)$$

In the second formula, the distortion $D(\mathbf{K}, \mathbf{T})$ is considered as the sum of local distortions occurring between two consecutive key frames as (5).

$$D(\mathbf{K}, \mathbf{T}) \equiv D_2(\mathbf{K}, \mathbf{T}) = \sum_{i=0}^N \left[\int_{k_i}^{t_i} d(t, k_i) dt + \int_{t_i}^{k_{i+1}} d(t, k_{i+1}) dt \right]. \quad (5)$$

Here, $k_0 = t_0$, $k_{N+1} = t_N$. This formula is useful in computing the breakpoints that minimize the distortion when the key frames are given in advance. A set of breakpoints that minimizes (5)

can be obtained as (6). Note again that depending on $d(\cdot)$ and frame feature vector, a closed-form solution like (2) can be obtained.

$$t_i = \arg \min_{k_i \leq t_i \leq k_{i+1}} \left[\int_{k_i}^{t_i} d(t, k_i) dt + \int_{t_i}^{k_{i+1}} d(t, k_{i+1}) dt \right], \quad i = 1, 2, \dots, N-1. \quad (6)$$

It is evident that for a specific pair of $\mathbf{K}$ and $\mathbf{T}$, $D_1(\mathbf{K}, \mathbf{T}) = D_2(\mathbf{K}, \mathbf{T})$. Once the distortion is defined as above, the key frame selection problem can be formulated as finding key frames and time-intervals that simultaneously minimize the distortion as (7).

$$\{\mathbf{K}^{\text{opt}}, \mathbf{T}^{\text{opt}}\} = \arg \min_{\mathbf{K}, \mathbf{T}} D(\mathbf{K}, \mathbf{T}). \quad (7)$$

3.2. An optimal approach

An optimal solution for (7) can be found by an exhaustive search method as follows. First, a set of pre-determined number of key frames is selected arbitrarily and using (6), non-overlapping time-intervals are obtained, which are illustrated in Fig. 2. Then, we compute the distortion using (3) or (5). If we wish to extract N key frames from M frames, $M C_N$ sets of key frames exist. Among them, the key frame set that generates minimum distortion is the optimal solution. This approach, however, may be impractical for a video shot containing large number of frames.

3.3. Iterative key frame selection

In this subsection, we propose an iteration-based method whose computational cost is significantly lower than the optimal method with slight degradation in the performance. First, note that when the positions of key frames (break-points) are selected independently with one another, their corresponding breakpoints (key frames) are computed automatically. For example, when a specific set $\mathbf{T}$ is determined in advance, we can easily obtain its corresponding $\mathbf{K}$ which minimizes (3) by collecting the key frames that minimize the local distortion in each of the local time-intervals using (4). In this case, however, it is obvious that even though the distortion is minimized in the sense of D_1 , it is not always minimal in the sense of D_2 . Namely, even if the key frames in $\mathbf{K}$ are optimal solutions in each of the time-intervals in $\mathbf{T}$, the breakpoints in $\mathbf{T}$ may not be optimal between two consecutive key frames in $\mathbf{K}$.

To solve this problem, we propose an iteration-based approach. Let $\mathbf{K}^{(m)} = \{k_1^{(m)}, k_2^{(m)}, \dots, k_N^{(m)}\}$ be a set of key frames, $\mathbf{T}^{(m)} = \{[t_0^{(m)}, t_1^{(m)}], [t_1^{(m)}, t_2^{(m)}], \dots, [t_{N-1}^{(m)}, t_N^{(m)}]\}$ be a set of time-intervals and $D^{(m)}$ be a distortion, all of which are obtained after m th iteration. Note again that $D^{(m)} = D_1(\mathbf{K}^{(m)}, \mathbf{T}^{(m)}) = D_2(\mathbf{K}^{(m)}, \mathbf{T}^{(m)})$. The overview of the proposed method is shown in Fig. 3.

First, the initial key frames $\mathbf{K}^{(0)}$ and initial time-intervals $\mathbf{T}^{(0)}$ are chosen based on the temporal variation. We select the pre-determined number of initial time-intervals and initial key frames so that

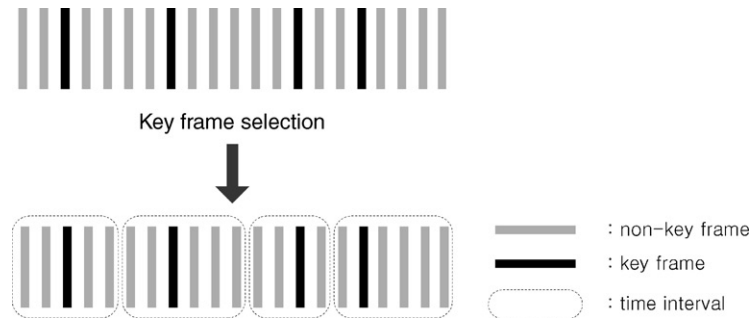


Fig. 2. An optimal key frame selection method.

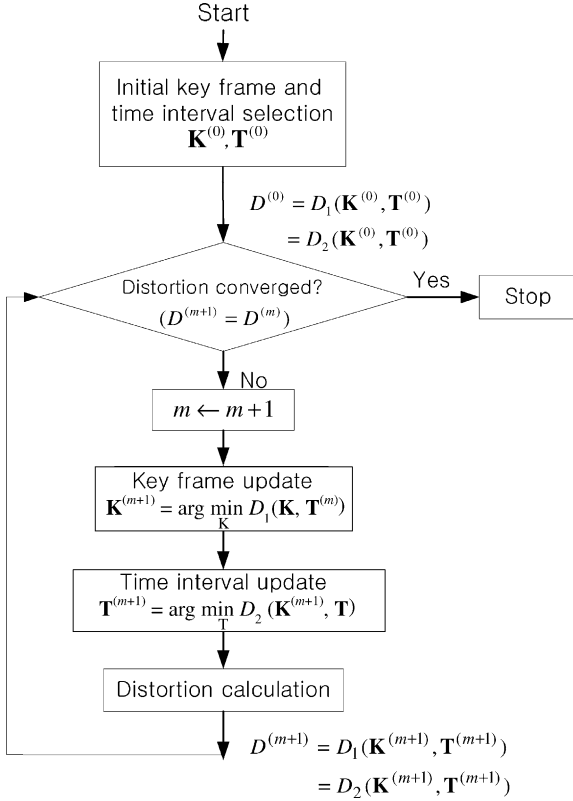


Fig. 3. Overview of the proposed method.

every initial time-interval may represent the same amount of temporal variation of visual content with one another, which will be explained in the next subsection. Then, we update the positions of key frames and breakpoints in turn. For example, after m th iteration, we update the positions of key frames as (8).

$$\mathbf{K}^{(m+1)} = \arg \min_{\mathbf{K}} D(\mathbf{K}, \mathbf{T}^{(m)}). \quad (8)$$

Here, $D(\mathbf{K}, \mathbf{T}) = D_1(\mathbf{K}, \mathbf{T})$ is used for distortion measure and (4) is used to compute the updated key frames. After updating key frames as above, the positions of breakpoints are updated as (9). In this case, $D(\mathbf{K}, \mathbf{T}) = D_2(\mathbf{K}, \mathbf{T})$ is used for distortion measure and (6) is used to compute the updated breakpoints.

$$\mathbf{T}^{(m+1)} = \arg \min_{\mathbf{T}} D(\mathbf{K}^{(m+1)}, \mathbf{T}). \quad (9)$$

We illustrate our method in case of two key frames in Fig. 4. Assume that after m th iteration, the two key frames $k_1^{(m)}$ and $k_2^{(m)}$ are given as shown in Fig. 4(a). In this case, using (6), a set of time-intervals $\mathbf{T}^{(m)}$ that minimizes (5) is obtained, which is shown in Fig. 4(b). The distortion obtained here, however, may not be minimal in the sense of D_1 . Namely, the key frames in Fig. 4(a) may not be optimal key frames in each time-interval of Fig. 4(b). Therefore, using (4), we adjust the key frames as in Fig. 4(c), obtaining a more reduced distortion. But again, in this case, the distortion may not be minimal in the sense of D_2 . Therefore, we adjust the time-intervals using (6) as Fig. 4(d). If we repeat the iteration until the distortion converges, we can obtain key frames and breakpoints which minimize the distortion in the sense of both D_1 and D_2 . After convergence, each key frame is locally optimal in its corresponding time-interval and, simultaneously, each breakpoint is locally optimal between two consecutive key frames.

Now, we prove that the distortion is guaranteed to converge as the iteration proceeds. First, after m th iteration, we adjust the positions of key frames in each time-interval using (4) in the sense of minimizing D_1 . Then, the following relation is always true for each of the local time-intervals.

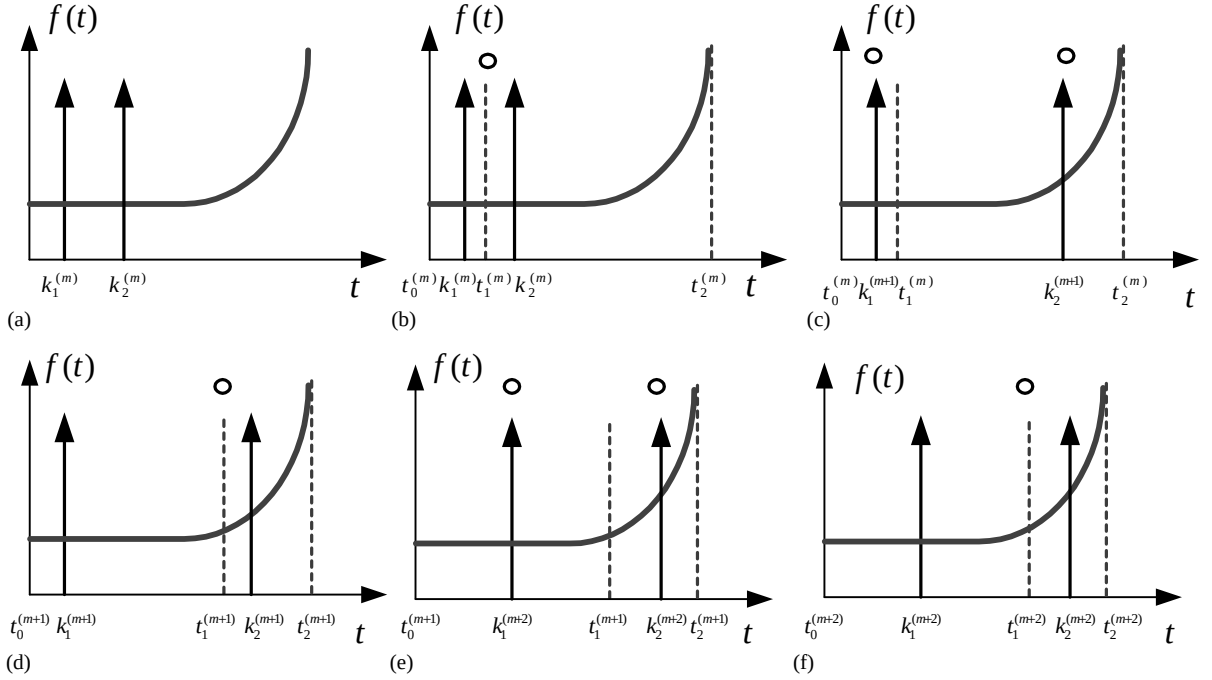
$$\int_{t_{i-1}^{(m)}}^{t_i^{(m)}} d(t, k_i^{(m+1)}) dt \leq \int_{t_{i-1}^{(m)}}^{t_i^{(m)}} d(t, k_i^{(m)}) dt, \quad i = 1, 2, \dots, N. \quad (10)$$

Namely, the following inequality is always true.

$$D_1(\mathbf{K}^{(m+1)}, \mathbf{T}^{(m)}) \leq D_1(\mathbf{K}^{(m)}, \mathbf{T}^{(m)}). \quad (11)$$

After updating the key frames as above, we can adjust the positions of breakpoints in the sense of minimizing D_2 using (6), obtaining the following relations:

$$\begin{aligned} & \int_{k_i^{(m+1)}}^{t_i^{(m+1)}} d(t, k_i^{(m+1)}) dt + \int_{t_i^{(m+1)}}^{k_{i+1}^{(m+1)}} d(t, k_{i+1}^{(m+1)}) dt \\ & \leq \int_{k_i^{(m+1)}}^{t_i^{(m)}} d(t, k_i^{(m+1)}) dt + \int_{t_i^{(m)}}^{k_{i+1}^{(m+1)}} d(t, k_{i+1}^{(m+1)}) dt, \quad i = 1, 2, \dots, N-1. \end{aligned} \quad (12)$$

Fig. 4. Example of iterative key frame selection (in case of $N = 2$).

Namely, the following inequality is always true.

$$D_2(\mathbf{K}^{(m+1)}, \mathbf{T}^{(m+1)}) \leq D_2(\mathbf{K}^{(m+1)}, \mathbf{T}^{(m)}). \quad (13)$$

Using (11), (13) and $D_1(\mathbf{K}^{(m+1)}, \mathbf{T}^{(m)}) = D_2(\mathbf{K}^{(m+1)}, \mathbf{T}^{(m)})$, we obtain the following inequality.

$$D_2(\mathbf{K}^{(m+1)}, \mathbf{T}^{(m+1)}) \leq D_1(\mathbf{K}^{(m)}, \mathbf{T}^{(m)}). \quad (14)$$

Then, from the fact that $D^{(m)} = D_1(\mathbf{K}^{(m)}, \mathbf{T}^{(m)})$ and $D^{(m+1)} = D_2(\mathbf{K}^{(m+1)}, \mathbf{T}^{(m+1)})$, the relation between current and previous distortion is obtained as (15), which indicates that the distortion is always reduced gradually as the iteration proceeds.

$$D^{(m+1)} \leq D^{(m)}. \quad (15)$$

It is noteworthy that Hanjalic's algorithm is not based on iteration, but based on the adjustments (trial-and-errors) of the first breakpoint. Therefore, it is evident that the direct comparison of our algorithm with Hanjalic's one in their computational complexity is meaningless and unfair. Of course, it is obvious that the computational complexity of our method for *one iteration* is

larger than that of Hanjalic's algorithm for *one step of adjustment* (i.e., for a particular value of first breakpoint) because Hanjalic's method has only to find the temporal location of key frames and breakpoints according to the (1) and (2), which have already been derived analytically. However, in Hanjalic's method, many steps of adjustments, the number of which cannot be properly estimated, are needed in order to accurately match the computed t_N with the actual length of a shot. On the contrary, it is evident that our algorithm is guaranteed to converge. Therefore, even though Hanjalic's algorithm seems to take much less time than our method for a particular value of the first breakpoint, the overall computational amount is not always less than our method.

3.4. Selection of initial key frames and time-intervals

In this subsection, the selection of initial key frames and initial time-intervals is explained. The

proposed sequential key frame selection is similar to non-uniform temporal sampling of 1-D signal. There may be various methods in selecting initial key frames and their corresponding time intervals. For example, as initial key frames, equi-distance sample frames or randomly selected sample frames may be chosen. However, in order to reduce the distortion defined in (3) or (5) as much as possible, it is desirable that key frames are selected more densely in the region where its visual content changes more rapidly. To achieve this, we select the initial time-intervals and their corresponding initial key frames so that the total magnitude of temporal variation within a video shot may be equally divided among them, namely, every initial key frame conveys the same amount of temporal visual content changes with one another. Taking this fact into consideration, we select the initial key frames and time-intervals as follows. First, we define the total temporal variation of visual content of a video shot as (16).

$$\Delta_V \equiv \int_0^M \left| \frac{df}{dt} \right| dt. \quad (16)$$

Then, a set of initial time-interval $\mathbf{T}^{(0)}$ is selected so that each of its elements may represent the same amount of temporal variation of visual content with one another as (17), which are illustrated in Fig. 5(b).

$$\begin{aligned} \int_{t_0^{(0)}}^{t_1^{(0)}} \left| \frac{df}{dt} \right| dt &= \int_{t_1^{(0)}}^{t_2^{(0)}} \left| \frac{df}{dt} \right| dt = \cdots = \int_{t_{N-1}^{(0)}}^{t_N^{(0)}} \left| \frac{df}{dt} \right| dt \\ &= \frac{\Delta_V}{N}. \end{aligned} \quad (17)$$

Lastly, in each initial time-interval, we select initial key frames using (4) as in Fig. 5(a). A more detailed explanation is found in [8].

4. Experimental results and discussions

Experiments are performed on a *foreman* sequence (0–398), which is a well-known MPEG video sequence. In this sequence, a foreman in front of a building talks with his head moving, followed by a fast camera panning near the 300th frame from the foreman to the construction-site.

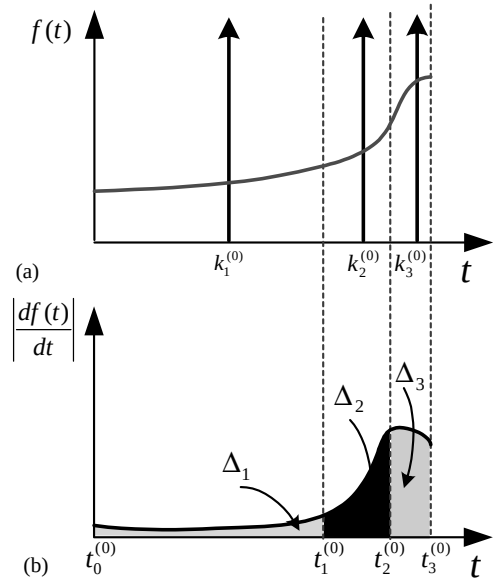


Fig. 5. Selection of initial time intervals and initial key frames (in case of $N = 3$). Here, $\Delta_1 = \Delta_2 = \Delta_3$.

Then, the sequence shows the construction site with the camera almost fixed. Therefore, we can say that it has three visual parts: foreman, panning and construction-site.

Fig. 6 shows several intensity-based visual features. Figs. 6(a) and (b) are the intensity variance and average of each frame, respectively. Fig. 6(c) is the histogram difference of consecutive frames. It can be observed that during the panning, there is a large variation in each of the frame feature vectors. Also note that when there is large variation in the intensity average or variance, the histogram difference is also large, which means that the intensity variance and average are good candidates for visual features. Of course, as discussed above, other features may be used depending on the user-interest, resulting in different set of key frames.

In the experiment, we compare the proposed method with the Hanjalic's method, which is one of the most quantitative and sophisticated sequential approaches, using the same visual feature and frame feature vector. For fair comparison with the Hanjalic's method, L_1 -norm is used to calculate the distance $d(t_i, t_j)$ between two frames and the cumulative absolute frame-by-frame feature difference is used as a frame feature vector $f(k)$ as

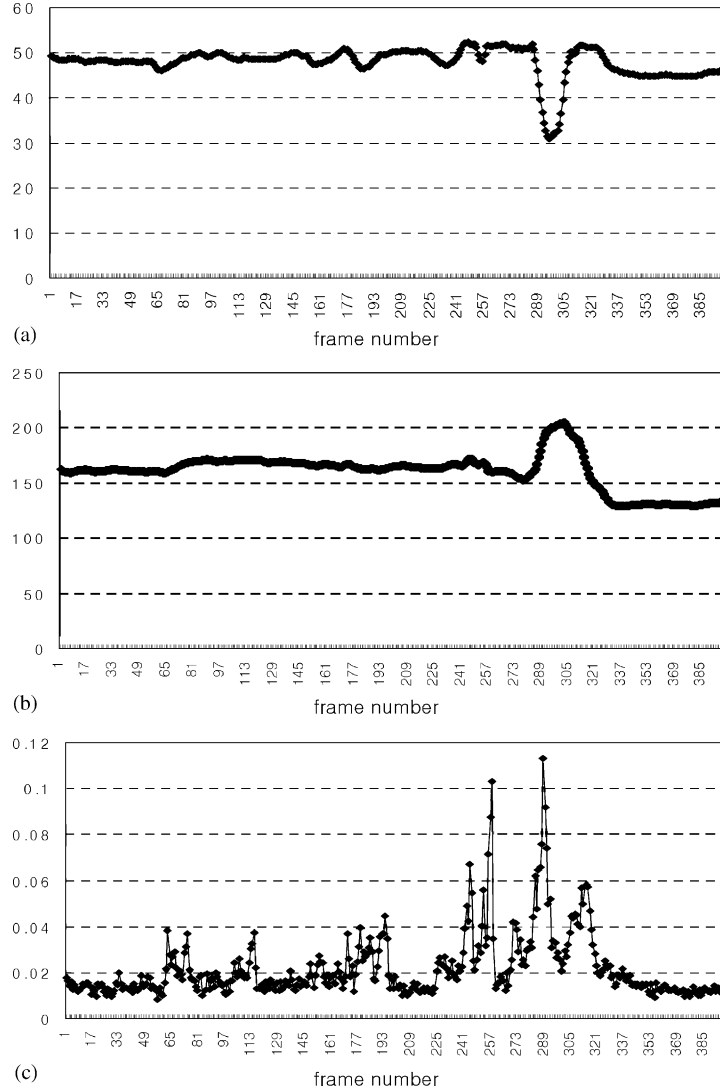


Fig. 6. Several frame feature vectors of *foreman* sequence: (a) variance of intensity, (b) average of intensity, (c) frame-by-frame histogram difference.

shown in (18). Here, $v(i)$ is the intensity variance at the i th frame.

$$f(k) = \sum_{i=1}^k |v(i) - v(i-1)|. \quad (18)$$

4.1. Objective performance comparison

For the objective performance comparison, we compare the two methods in the sense of rate-

distortion. Here, *rate* means ratio of the number of assigned key frames to that of total frames. Fig. 7(a) demonstrates the performance. Here, the x -axis and y -axis represent the pre-determined number of key frames and the distortion, respectively. From this figure, it is observed that our method is superior to the Hanjalic's method in the rate-distortion sense. Note that, as discussed above, both of the two methods try to make key frames and breakpoints locally optimal. But in the

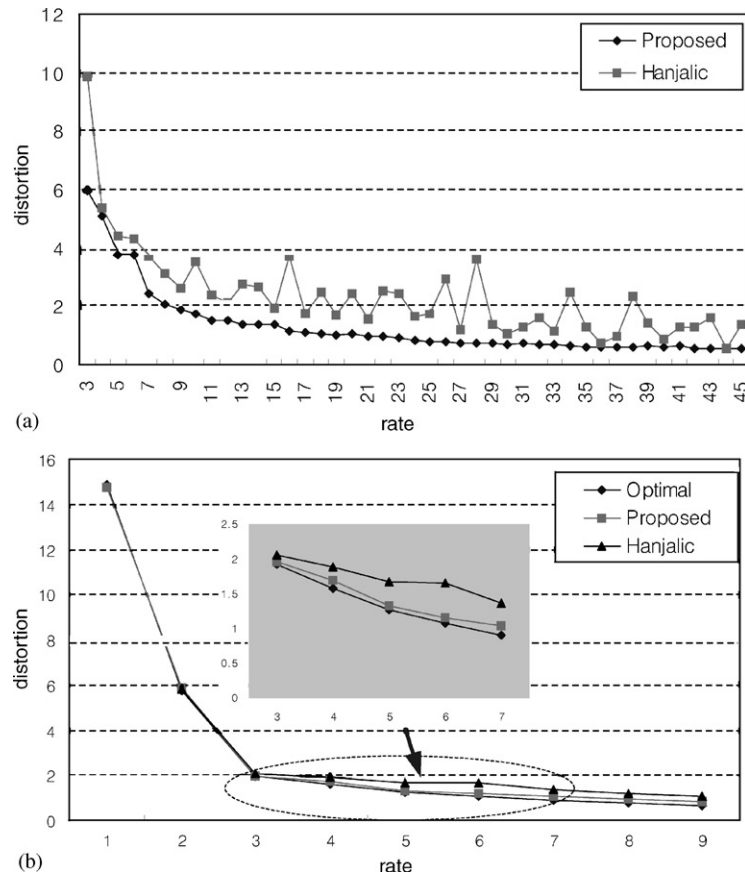


Fig. 7. Experimental result: performance comparison in the sense of rate distortion.

Hanjalic's method, although the distortion tends to decrease as the number of extracted key frames increases, some distortion values are fluctuating. This indicates that the solution found by Hanjalic's method is not always global minimum but sometimes, local minimum, local maximum or even global maximum in the worst case. On the contrary, since our approach is based on the iteration that is guaranteed to converge, it does not show such undesired fluctuations. Moreover, in the Hanjalic's method, the current key frame is selected from the recursive relation with the previous key frame and breakpoint. To satisfy this recursive relation, some key frames are localized at undesired positions, which will be shown later. Due to these facts, Hanjalic's method shows poorer performance than our approach.

Next, we compare the proposed method with the optimal method discussed in Section 3.2 in the rate-distortion sense. In this case, however, it is computationally unfeasible to obtain a rate-distortion curve of the optimal method with the original 399 frames as Fig. 7(a). For example, to compare ${}_{399}C_5$ sets of key frames, we find that it takes about 95 days with Pentium-III (800MHz CPU) personal computer. Therefore, we use only 50 consecutive frames (frame 271 frame 320) collected from a *foreman* sequence and evaluate the rate-distortion characteristics using these frames. It contains all of the three visual parts. We compare the proposed method with the optimal method, together with the Hanjalic's method again. In Fig. 7(b), it is shown that the proposed method achieves almost the same

performance with the optimal method, while reducing the computational complexity greatly.

4.2. Subjective performance comparison

In this subsection, we compare the selected key frames subjectively. Figs. 8(a) and (b) show ten key frames selected by the proposed method and Hanjalic's method, respectively, from the original *foreman* sequence. Note that the key frames selected by Hanjalic's method represent the foreman part well. These key frames, however, cannot describe the camera panning part well. In the panning part, it selects only one key frame (305th frame) although large temporal variation exists during the panning. In the construction-site part, too many key frames are selected compared with its duration time and content variation. On the other hand, as illustrated in Fig. 8(a), the key

frames selected by the proposed method represent the total shot well. In the foreman part, since it takes relatively large portions among all frames, five key frames are selected. In the panning part, which has relatively short duration but large variation, four key frames are selected. As for the construction-site part, since this part has almost no variation and short duration, only one key frame is selected.

Fig. 9 shows the temporal distribution of the extracted key frames. As Fig. 9(a) shows, the proposed method selects key frames almost uniformly in the foreman part and very densely in the panning part. Also, note that our method takes not only the temporal variation but also its duration time into account, which indicates that the selected key frames represent the temporal variation of a shot globally well. Meanwhile, as illustrated in Fig. 9(b), Hanjalic's method cannot



Fig. 8. Experimental result. Selected key frames of (a) the proposed method and (b) Hanjalic's method.

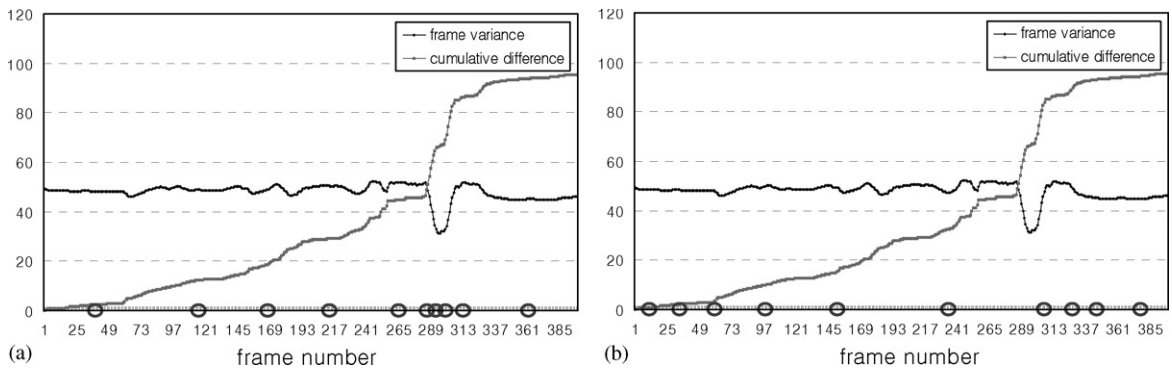


Fig. 9. Distribution of key frames along a video shot (small circles on the x-axis indicate the temporal positions of key frame): (a) proposed method and (b) Hanjalic's method.

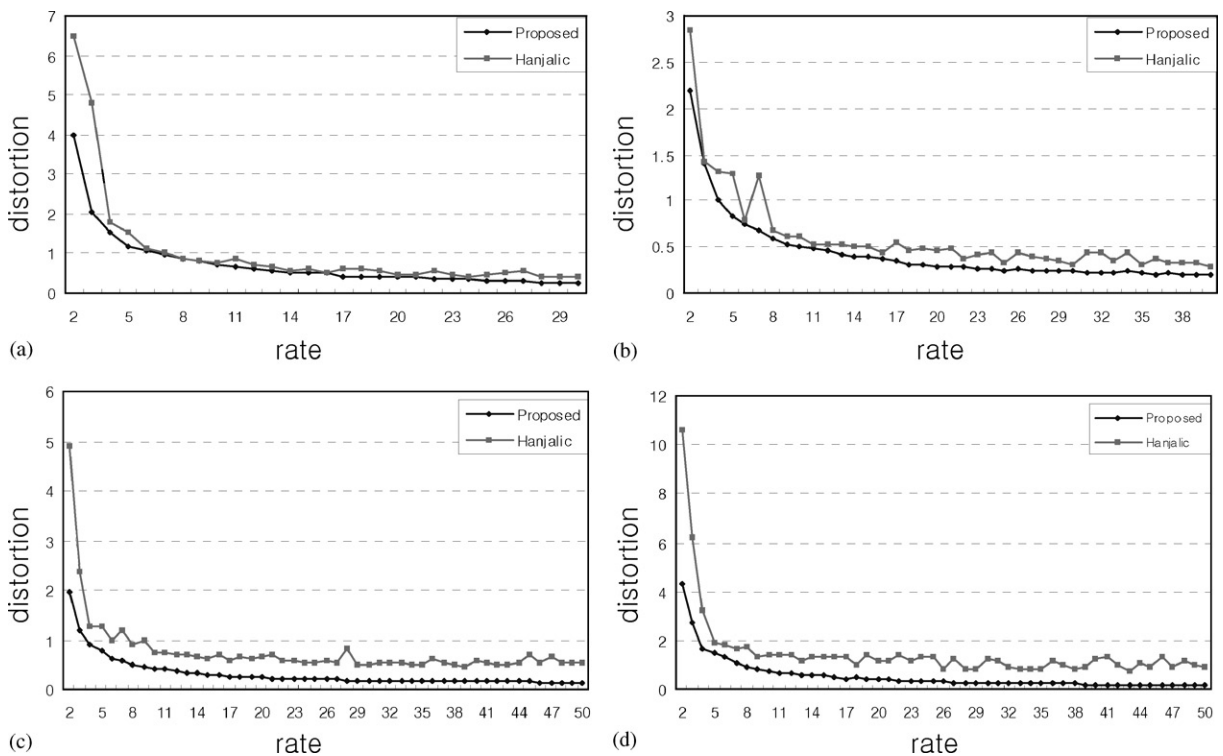


Fig. 10. Comparison of rate-distortion performance for (a) *bike* sequence, (b) *home-made outdoor* sequence, (c) *coast* sequence, (d) *table tennis* sequence.

represent the panning part well. As discussed just before, to satisfy this recursive relation, some key frames are localized at undesired positions. The seventh key frame (frame 305) in the Hanjalic's method is a typical example. To represent the panning part, a more previous frame should be

selected instead of this frame. But Hanjalic's method selects the 305th frame as a key frame just to satisfy the recursive relation.

Lastly, in order to show that the proposed algorithm outperforms Hanjalic's method for a wide range of sequences, we perform the same

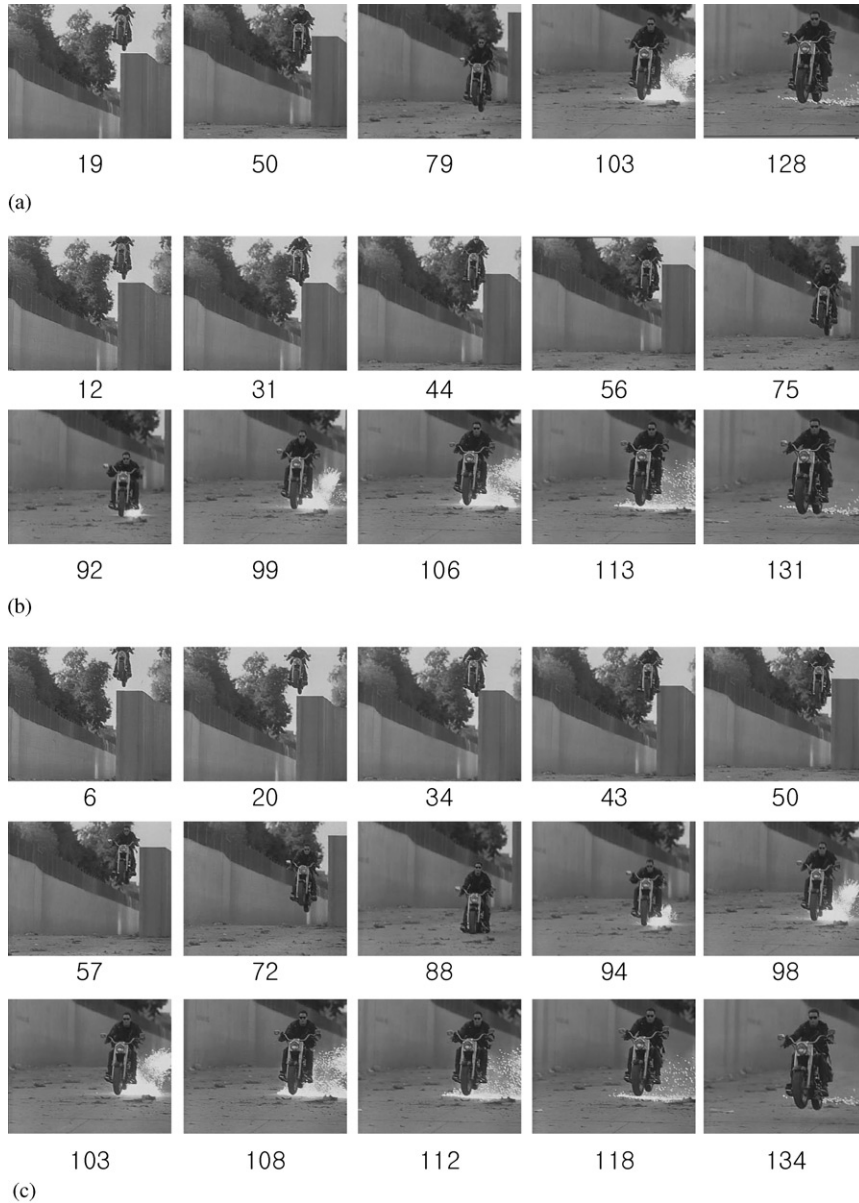


Fig. 11. Selected key frames for *bike* sequence by the proposed method (a) $N = 5$, (b) $N = 10$, (c) $N = 15$.

experiment on a *bike* sequence (0–148), a home-made outdoor sequence (0–196), *coast* sequence (1–300), and *table tennis* sequence (1–300). Among them, *coast* and *table tennis* are widely and well known MPEG test sequences. The rate-distortion characteristics are shown in Fig. 10, which shows that our algorithm is better than Hanjalic's

method in the rate-distortion sense for a wide range of sequences. Also, in order to show that the resulting key frames are appealing to human observers, we select key frames for various N from a *bike* sequence. The result is shown in Fig. 11. As can be seen, the selected key frames can summarize the visual content of a sequence well.

4.3. Effect of initial key frames

To show the effect of initial key frames on the performance, we compute the rate-distortion curve and the average iteration number for the three initial key frame selection methods mentioned in Section 3.4; proposed visual content-adaptive selection, uniform-sample selection and random-sample selection. As shown in Fig. 12, although the random-sample selection shows the poorest performance for a small number of key frames, the rate-distortion performance is almost the same for the three methods if the number of key frames is larger than a certain value. In case of random-sample selection with a small number of key frames, the initial key frames may be put in one side, resulting in poorer rate-distortion performance. Nevertheless, if the number of key frames increases, the initial key frames are selected more uniformly even in case of random-sample selection. Meanwhile, the uniform-sample selection shows almost the same rate-distortion performance as the proposed one since the initial key frames are not crowded into one side. From this fact, we conclude that if only the initial key frames are distributed along the entire shot uniformly within a certain level, the resultant key frames are converged to almost identical image frames

irrespective of the initial key frame selection method.

However, when the proposed initial key frame selection method is used, the average iteration number is less than any other methods as shown in Table 1. The average iteration number is more reduced in case of the proposed method because the proposed initial key frames are nearer to the converged key frames than those selected by other methods. This means that by selecting initial key frames based on the visual content, the computation time is more reduced than any other methods.

5. Conclusions and further works

In this paper, we present a sequential key frame selection method when the number of key frames

Table 1
Performance comparison of initial key frame selection methods

	Proposed	Uniform sampling	Random selection
Average iteration number	4.98	5.12	7.58

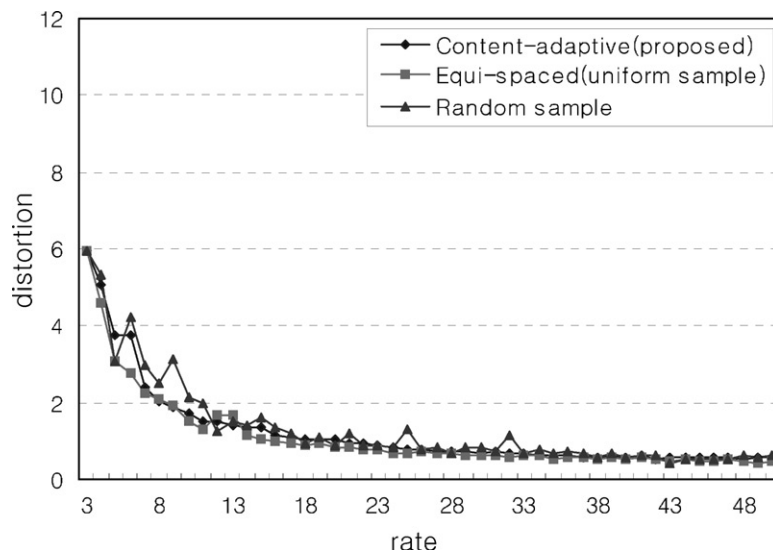


Fig. 12. Effects of initial key frames on the rate-distortion performance. In this experiment, *foreman* sequence is used.

is given as a constraint. It first selects the pre-determined number of initial key frames and initial time-intervals based on the temporal variation of visual content. Then, it adjusts the positions of key frames and time-intervals by iteration, which reduces the distortion gradually. The proposed method can select the assigned number of key frames without any threshold or first breakpoint specification and irrespective of the type of video sequence. In addition, although Hanjalic's method is only applicable to the cumulative (non-decreasing) frame feature vector, our method can be used not only in case of non-decreasing frame feature vector, but also in the case of any type of frame feature vector. Experimental results demonstrate the improved performance of our algorithm over the existing approach both in the R-D sense and subjective measure.

The key frames selected by our method can be used in computing the visual similarity between two video segments more reliably [13,14] since they can approximate the temporal variations better than those selected by other existing approaches. Also, our method can be easily extended to the distortion-constrained key frame selection scheme by adjusting the number of extracted key frames. In the future, we are going to perform research on the efficient video retrieval scheme using the selected key frames.

References

- [1] R. Brunelli, O. Mich, C.M. Modena, A survey of the automatic indexing of video data, *J. Visual Commun. Image Representation* 10 (1999) 78–112.
- [2] H.S. Chang, S. Sull, S.-U. Lee, Efficient video indexing scheme for content-based retrieval, *IEEE Trans. Circuits Syst. Video Technol.* 9 (8) (1999) 1269–1279.
- [3] N.D. Doulamis, A.D. Doulamis, Y.S. Avrithis, S.D. Kollias, Video content representation using optimal extraction of frames and scene, in: *Proc. ICIP 98*, Vol. 1, 1998, pp. 875–879.
- [4] U. Gargi, R. Kasturi, S.H. Strayer, Performance characterization of video-shot-change detection methods, *IEEE Trans. Circuits Systems Video Technol.* 10 (1) (2000) 1–13.
- [5] A. Hanjalic, R.L. Lagendijk, J. Biemond, A new method for key frame based video content representation, in: *Proceedings of the First International Workshop on Image Databases and Multi-Media Search*, Amsterdam (NL), 1996.
- [6] F. Idris, S. Panchanatha, Review of image and video indexing, *J. Visual Commun. Image Representation* 8 (2) (1997) 146–166.
- [7] R.L. Lagendijk, A. Hanjalic, M. Ceccarelli, M. Soletic, E. Persoon, Visual search in a SMASH system, in: *Proc. ICIP 96*, Vol. 3, 1996, pp. 671–674.
- [8] H.-C. Lee, S.-D. Kim, Rate-driven key frame selection using temporal variation of visual content, *IEE Electron. Lett.* 38 (5) (February 2002) 217–218.
- [9] H.-C. Lee, C.-W. Lee, S.-D. Kim, Abrupt shot change detection using an unsupervised clustering of multiple features, *Proc. ICASSP 4* (2000) 2015–2018.
- [10] MPEG7: Context, Objectives and Technical Roadmap, V.11, ISO/IEC JTC1/SC29/WG11/N2729, March 1999, Seoul.
- [11] A. Nagasaka, Y. Tanaka, Automatic video indexing and full-video search for object appearances, in: *Proc. of 2nd Working Conf. on Visual Database Systems*, October 1991, pp. 119–133.
- [12] B. Shahrar, D.C. Gibbon, Automatic generation of pictorial transcript of video programs, in: *Proceedings of SPIE Digital Video Compression: Algorithms and Technologies*, San Jose, CA, 1995, pp. 512–519.
- [13] M.-K. Shan, S.-Y. Lee, Content-based video retrieval based on similarity of frame sequence, in: *Proceedings of the International Workshop on Multimedia Database Management Systems*, 1998, pp. 90–97.
- [14] Y.-P. Tan, S.R. Kulkarni, P.J. Ramadge, A framework for measuring video similarity and its application to video query by example, in: *Proceedings of the International Conference on Image Processing*, 1999.
- [15] M.M. Yeung, B. Liu, Efficient matching and clustering of video shots, in: *Proc. ICIP 95*, Vol. 1, 1995, pp. 338–342.
- [16] M.M. Yeung, B.-L. Yeo, Video visualization for compact presentation and fast browsing of pictorial content, *IEEE Trans. Circuits Syst. Video Technol.* 7 (5) (1997) 771–785.
- [17] M.M. Yeung, B.-L. Yeo, Segmentation of video by clustering and graph analysis, *Comput. Vision Image Understanding* 71 (1) (1998) 94–109.
- [18] H.J. Zhang, J. Wu, D. Zhong, S.W. Smoliar, An integrated system for content-based video retrieval and browsing, *Pattern Recognition* 30 (4) (1997) 643–658.
- [19] Y. Zhuang, Y. Rui, T.S. Huang, S. Mehrotra, Adaptive key frame extraction using unsupervised clustering, in: *Proc. ICIP*, Vol. 1, 1998, pp. 866–870.