Scalable Hierarchical Video Summary and Search

Sanghoon Sull [*a], Jung-Rim Kim [a], Yunam Kim [a], Hyun Sung Chang [b], and Sang Uk Lee [c]

[a] School of Electrical Engineering, Korea University, Seoul, Korea

[b] Electronics and Telecommunications Research Institute, Taejon, Korea

[c] School of Electrical Engineering, Seoul National University, Seoul, Korea

ABSTRACT

Recently, a huge amount of the video data available in the digital form has given users to allow more ubiquitous access to visual information than ever. To efficiently manage such huge amount of video data, we need such tools as video summarization and search. In this paper, we propose a novel scheme allowing for both scalable hierarchical video summary and efficient retrieval by introducing a notion of fidelity. The notion of fidelity in the tree-structured key frame hierarchy describes how well the key frames at one level are represented by the parent key frame, relative to the other children of the parent. The experimental results demonstrate the feasibility of our scheme.

Keywords : scalable hierarchical video summary, content-based video retrieval/search, video indexing, fidelity

1. INTRODUCTION

Nowadays, as the network bandwidth is rapidly increasing, a huge number of video streams are widely available through the Internet. The tremendous increase of the amount of multimedia data, however, bears the problems on the managing of archives, and the search and browsing of preferable video streams. Thus, there appears a strong need for efficiently indexing video data allowing for summary and search.

Recently, several researchers [2-4] have developed methods of summarizing the input video. Arthur Pope, et al. [2] proposed a scheme for summarizing video content using mosaics and moving object trajectories. Within each video clip, static scene content is summarized by a mosaic, and dynamic content, by segmenting and tracking moving objects. The method was designed to exploit the location and time information of aerial reconnaissance video that has some noteworthy characteristics for the retrieval and visualization. Peter J.Macer and Peter J.Thomas [3] presented another approach to the video summary. They have proposed a method in which a representative single frame per each shot is automatically selected from a given video sequence to form a storyboard. The storyboard is a representation used in the production of all types of film and television production, and consists of a series of still images each of which summarizes the scene and action in each shot. Shingo Uchihashi and Jonathan Foote [4] proposed a video summarizing scheme using shot importance. The importance measure of each shot was assumed to be larger if the shot is long, and smaller if the cluster weight is large.

Many researchers have also developed video search methods based on key frame hierarchy [1, 5], and matching and clustering of shots [6]. The approaches based on the matching with key frames [1, 5] construct an efficient key frame hierarchy, resulting in much less number of frame comparisons than that required when the serial search is applied to the whole set of key frames. For example, H. S. Chang, et al. [1] presented a set-theoretic key frame extraction method and its application to video search.

In this paper, we present a novel scheme allowing for both scalable hierarchical video summary and efficient retrieval by introducing the notion of fidelity. This scheme is an improved version of the tree-structured key frame hierarchy proposed in [1]. The tree-structured key frame hierarchy presented in [1] was constructed by the hierarchical application of combinatorial extraction of (sub)minimal key frames. It was demonstrated that the resulting tree-structured key frame hierarchy, based on the branch-and-bound scheme, greatly reduces the number of frame comparisons required for the

[*] Correspondence: Email : sull@mail.korea.ac.kr; WWW : http://mpeg.korea.ac.kr/~sull

retrieval. The improved tree-structured key frame hierarchy based on the notion of the fidelity in this paper further reduces the number of frame comparisons for the retrieval and also allows for the scalable video summary at the same time. The notion of fidelity in the tree-structured key frame hierarchy describes how well the key frames at one level are represented by the parent key frame, relative to the other children of the parent. The functionality of the scalable video summary allows for selecting an arbitrary number of key frames specified by a user that best represent the whole video. For example, suppose that a user wants to see the 1 minute-summary of the 2-hour video. Then, the 1800 key frames need to be selected from the key frame hierarchy that covers the 216,000 original frames.

The paper is organized as follows. Section 2 describes a tree-structured key frame hierarchy. Section 3 presents the use of its hierarchy for scalable hierarchical video summary and search, and Section 4 shows the experimental results for video summary and search. Finally, Section 5 concludes the paper.

2. TREE-STRUCTURED KEY FRAME HIERARCHY

In this section, we describe a new tree-structured key frame hierarchy. We first present the feature vector used. Then, we briefly describe a key frame extraction method and its hierarchical application to construct a tree-structured key frame hierarchy with fidelity value attached to each node [1]. Then, we present an improved key frame hierarchy with fidelity value on each edge in the tree.

2.1. Feature Vector for a Frame

For the content-based retrieval and summary, the extraction of a feature vector from each frame in the input video is first performed. There are variety of features such as color, texture and shape. In our current implementation, we use a color correlogram that expresses how the spatial correlation of pairs of colors changes with distance in an image [7].

2.2. Key Frame Extraction

The use of a good key frame extraction method is also important for efficient video summary and search. Tonomura et al. [8] used the first frame of each shot as a key frame. This system can also provide an alternate representation of a video sequence that uses key frames that are evenly spaced, ignoring shot boundaries. Nagasaka et al. [9] also used the first frame of each shot as the shot's key frame. Ueda et al. [10] represent shots with two key frames, the first and last frames of each shot. Ferman et al. [11] use clustering on the frames within each shot. The frame closest to the center of the largest cluster is selected as the key frame for that shot. And Taniguchi et al. [12] generate a composite image to represent shots with camera motion. Yueting Zhuang et al. [13] extract key frames based on the visual content complexity indicator. Wayne Wolf [14] used the shot motion indicator to extract key frames. Wolf first computed the optical flow for each frame, and then computed a simple motion metric based on the optical flow. Also, he analyzed the metric as a function of time to select key frames at the local minima of motion. H. S. Chang, et al. [1] presented a set-theoretic key frame extraction method to find a compact set of key frames that can represent a video segment for a given degree of fidelity. In this paper, we use the approach in [1] since its hierarchical application yields a key frame tree hierarchy containing the information on how well the parent key frame represents the children.

2.3. Tree-Structured Key Frame Hierarchy with Fidelity in Each Node

By considering the whole set of key frames extracted from all shots, a key frame tree hierarchy is constructed by clustering them in a bottom-up way. In [1], the key frame extraction problem is modeled as choosing a compact set of samples (key frames) among many data points (frames in a video shot), while keeping the distortion less than a given threshold, which is analogous to the vector quantization scheme. The combinatorial selection of a (sub)minimal set of key frames under the given fidelity constraint yields (sub)optimal results in terms of rate-distortion (R-D) performance. By using the combinatorial property, the extraction method [1] is hierarchically applied to higher levels, starting from the frames in each shot, yielding a tree-structured key frame hierarchy. The key frame hierarchy is a multilevel abstract of a video, in which each level represents the whole video content at different level of details.

Consider Fig. 1 that represents a key frame hierarchy where the fidelity value is attached to each node [1] (we will call this fidelity as the node fidelity). Thus, the node A represents its three subtrees rooted at the nodes B, C, D, respectively within an extent of the fidelity value f_A. This value is computed by an encoder, for example by computing dissimilarities between A and all of its children, selecting the maximum, and then normalizing it after taking the reciprocal of the maximum. Thus, the fidelity value 1 means that the node A perfectly represents all of its children.

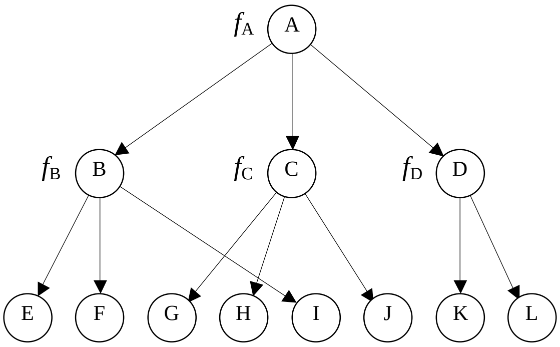


Fig.1. An example of the key frame hierarchy proposed in [1].

2.4. Improved Tree-Structured Key Frame Hierarchy with Fidelity on Each Edge

In this subsection, we propose an improved key frame hierarchy where the fidelity values are attached to each edge. Compared to the fidelity in each node, the fidelity on each edge not only increases the video search efficiency but also provides a scalable hierarchical video summary. The detailed description of its applications to search and summary will be presented in the next section.

Suppose now that we want to select two nodes that best represent the whole nodes. In other words, we need to split one edge in the original tree, resulting in two separate trees whose roots now represent their own subtree, respectively. Then, one of the possible ways of defining how well the two selected nodes represent all of the nodes is the minimum of the fidelity values corresponding to the two trees. After selecting the root A, we need to decide which node to select. One of the reasonable ways is to select one from the nodes B, C and D whose fidelity value is the minimum after splitting. Thus, if the fidelity value is defined in a node, a decoder needs to recompute the fidelity value after splitting. Since this is not efficient and further the decoder does not know the dissimilarity measure that the encoder used, there possibly occurs a problem.

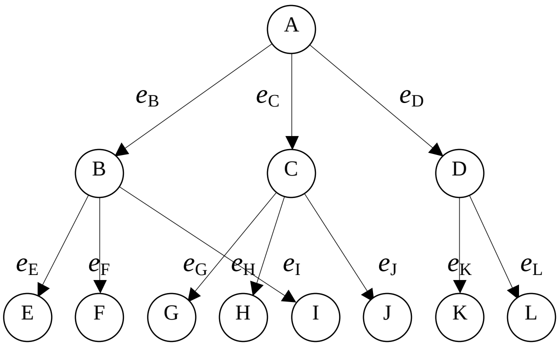


Fig. 2. An example of the improved key frame hierarchy with fidelity on edge.

Therefore, we propose to move the fidelity value from a node to an edge in the video hierarchy structure as shown in Fig. 2 (this fidelity will be called as the edge fidelity). In our current implementation, the fidelity value e_S of the parent frame p for one of its subtrees S, is computed as follows:

$$e_S = 1 - \max(d(p, s_i)), \quad for\ \forall s_i \in S, \tag{1}$$

where $d()$ denotes normalized distance from 0 to 1. For example, in Fig 2. the e_B is computed as

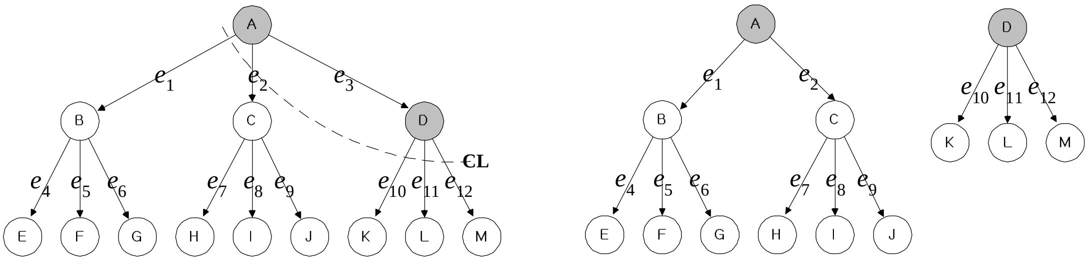
$$e_B = 1 - \max(d(A,B), d(A,E), d(A,F), d(A,I)).$$

3. APPLICATIONS TO SUMMARY AND SEARCH

In this section, we present two important applications of the improved tree structure with the fidelity attached to each edge, the scalable video summary and search. As described before, the fidelity value describes the degree of how well a given key frame represents its children and thus it provides useful information on the goodness of a key frame.

3.1. Scalable Hierarchical Video Summary

Suppose that we want to find the N numbers of key frames to best represent the whole video. One of the systematic way is to select the N key frames resulting in the maximum fidelity, which is illustrated below.



(a) (b)

Fig. 3. An example of the key frame hierarchy.

The following notations will be used to denote the components in the key frame hierarchy.

Table 1. Preliminary notation

notation	Expression	Meaning
T	whole tree	
T_α	subtree rooted at node α	
$e_i =< \alpha, \beta >$	edge from node α to node β	Fidelity with which the node α represents T_β
E_α	$\mathrm{E}_\alpha = \{e_i \mid e_i =< \alpha, x >, x \in \mathrm{T}\}$,	set of edges whose starting end is the node α
f_α	$f_\alpha = \min_{e_i \in \mathrm{E}_\alpha} e_i$	Fidelity with which the node α represents T_α
K	selected set of key frames	
f_{K}		Fidelity with which K represents T

First, let us consider the case of *N = 1*. It is natural that {A} should be K. The fidelity f_{K} is calculated by

$$f_{\mathrm{K}} = f_A = \min(e_1, e_2, e_3). \tag{2}$$

Now, we are encountered with the case of *N = 2*. There are three possible choices for K, which are listed in Table 2.

Table 2. Possible schemes for *N = 2*

scheme	K	Fidelity, f_{K}
(a)	{A, B}	$\min(e_2, e_3, e_4, e_5, e_6).$
(b)	{A, C}	$\min(e_1, e_3, e_7, e_8, e_9).$
(c)	{A, D}	$\min(e_1, e_2, e_{10}, e_{11}, e_{12}).$

Among them, for example, consider the case of (c), as illustrated in Fig. 3. This choice decomposes the original tree in Fig. 3(a) into two subtrees in Fig. 3(b). Notice that the decomposition may make changes in the degrees and/or fidelity. (We define the degree of a node as the number of subtrees which it covers.) For example, in the original tree of Fig. 3(a), the degree of A is 3, while 2 in the decomposed tree of Fig. 3(b). The fidelity f_{K} is calculated by

$$f_{\mathrm{K}} = \min(f_A, f_D) = \min(\min(e_1, e_2), \min(e_{10}, e_{11}, e_{12})) = \min(e_1, e_2, e_{10}, e_{11}, e_{12}). \tag{3}$$

Now, our aim is to find the *max-cut*, which maximizes the minimum edge cost cutted by CL in the tree of Fig. 3(a). The *max-cut* finding algorithm is depicted in Fig. 4. In this way, we can systematically select the N key frames to best represent the whole video.

```
add root_node to K;
while ( $card$(K) < $N$ ) {
    let <$\alpha$,$\beta$> be a least cost edge such that $\alpha \in$ K and $\beta \notin$ K;
    add $\beta$ to K;
}
```

Fig. 4. The *max-cut* finding algorithm.

3.2. Video Search

For the efficient hierarchical search, it is desirable to use some guidance information. Suppose a key frame hierarchy with the edge fidelity is constructed from a video. If a user has a query image and wants to find all the relevant frames in a video sequence, then the depth-first-search with pruning at each node can carry out the desired functionality much faster than the serial search. The pruning at each node in the tree occurs whenever the key frame corresponding to the node is not similar to the query image and the fidelity value of the edge is high. Thus, we can achieve the fast access to the frames relevant to a query image in a video sequence or quick rejection for the video sequences whose contents are irrelevant to the query.

The video search problem based on key frames is to find all the relevant key frames whose metric distance from a given query image is less than d_0 given by a user. The use of fidelity values in a key frame hierarchy increases the search efficiency, which is illustrated with reference to Fig. 3(a) as follows: From the fidelity values of the edges connected to key frame A, we can compute the distance d_i between A and its children by $d_i = 1 - e_i$, (i = 1,2,3). For simplicity, assume $d_1 < d_2 < d_3$. In traversing the nodes in the tree-structured key frame hierarchy to find the relevant key frames, the pruning of a subtree rooted at B occurs when $d > d_1 + d_0$ where d denotes the distance (or dissimilarity) between key frame A and query image I_Q. In Table 3, the pruning conditions for two schemes are compared where we can see that the proposed notion of fidelity allows better search performance than the node fidelity [1].

Table 3. Comparison of the pruning conditions of two schemes.

Search Schemes Condition	Edge Fidelit (Proposed)	Node Fidelity (Previous [1])
$d < d_1 + d_0$	No pruning	No pruning
$d > d_1 + d_0$ && $d < d_2 + d_0$	Pruning of the subtree rooted at B	No pruning
$d > d_2 + d_0$ && $d < d_3 + d_0$	Pruning of the subtrees rooted at B and C	No pruning
$d > d_3 + d_0$	Pruning all subtrees	Pruning all subtrees

4. EXPERIMENTAL RESULTS

We have implemented a prototype system for video summary and search as shown in Fig. 5. Our system is based on server-client architecture. For our experiments we used color correlogram as the feature vector and the key frame extraction method in [1], to construct a 5-level key frame hierarchy using bottom-up procedure. We tested our approach using two 10-minute test videos contributed to MPEG-7 and a 1-minute movie clip.

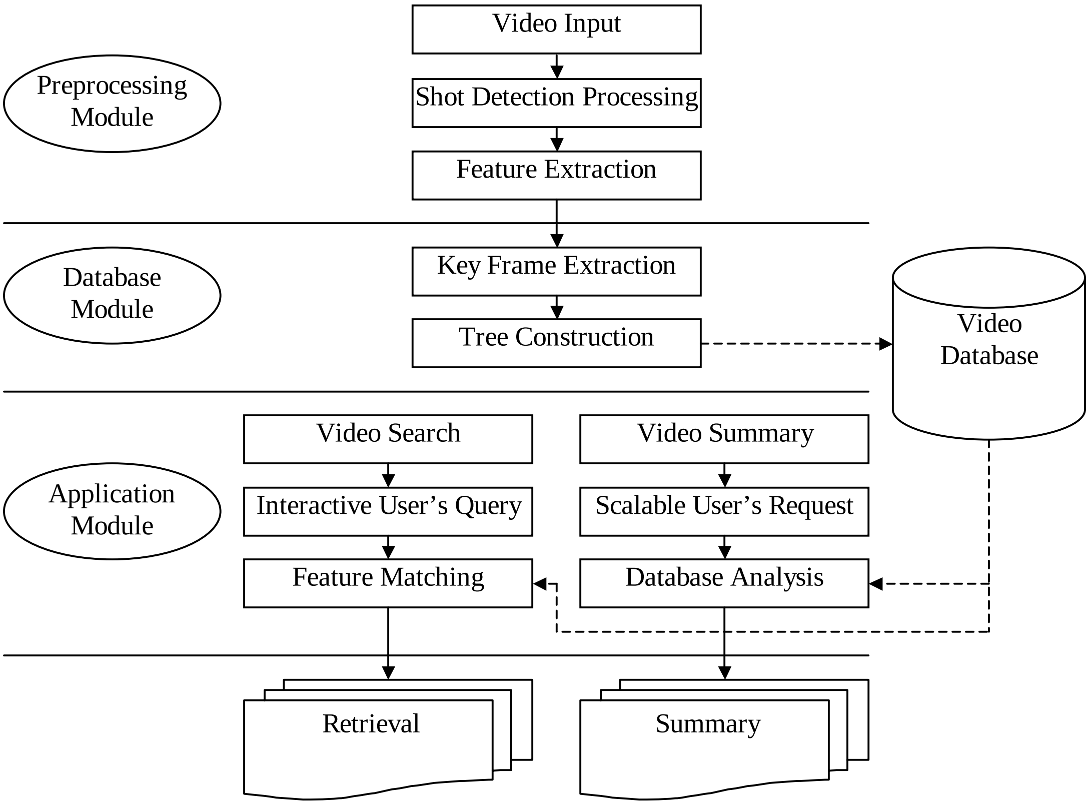
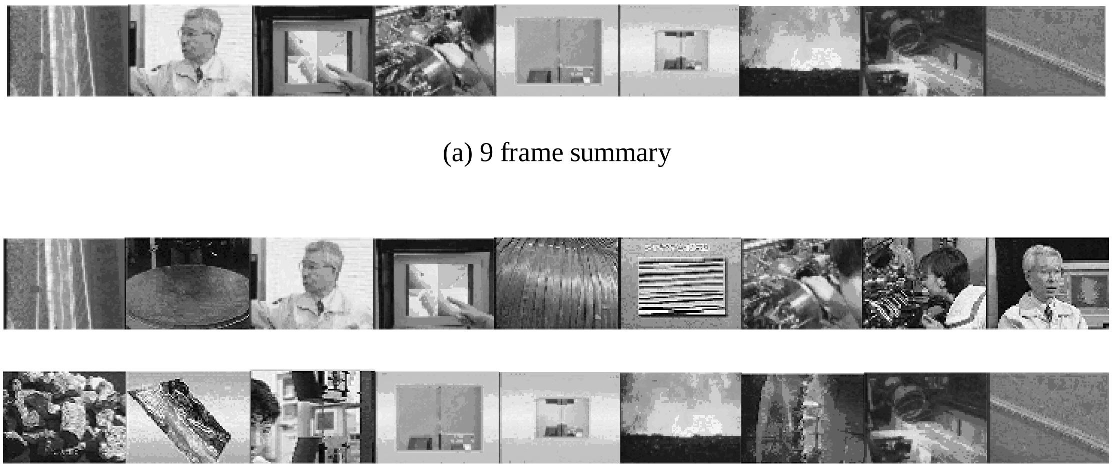


Fig. 5. Our prototype system for video summary and search.

4.1. Result for Scalable Hierarchical Video Summary

We have tested our video summary algorithm using the NHK documentary, MPEG-7 V10 whose length is about 10 minutes. Fig. 6 shows the result of scalable video summary. Although we have shown two summary results, the proposed scheme can provide a scalable summary using an arbitrary number of key frames depending upon the network bandwidth and user's preference.



(a) 9 frame summary

(b) 18 frame summary

Fig. 6. Result of scalable video summary.

4.2. Result for Video Search

We measure the improved ratio of search speed as

$$Improved\ ratio = \frac{N_n - N_e}{N_n} \times 100(\%), \tag{4}$$

where N_n is the number of image comparisons when using the node fidelity and N_e is the one when using the edge fidelity.

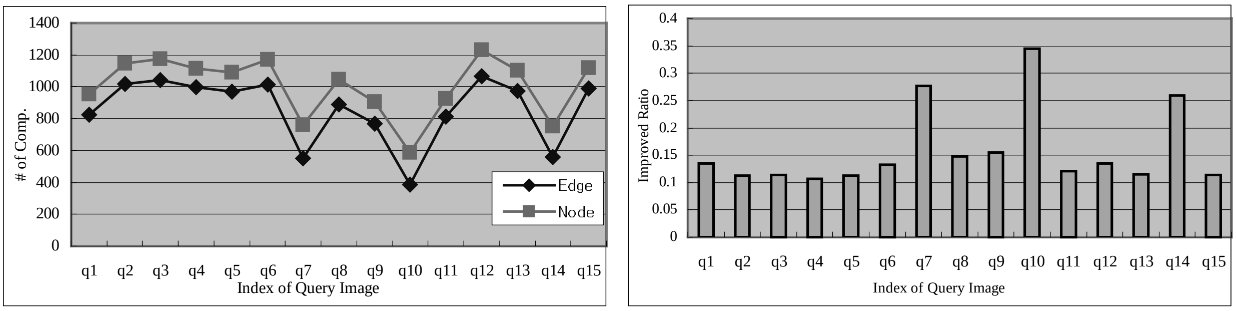
Table 4 shows the search result for the given set of query images to each video stream. The speed of the search using edge fidelity is about 11.6% better than that of the search using the node fidelity and about 33.3% better than that of the serial search.

Table 4. Search result for each video stream (d_0 = 0.3).

Video	Movie clip	Sports (MPEG-7 V18)	Documentary (MPEG-7 V10)
Length(min:sec)	01:00	09:04	10:04
# of frames	1453	18111	16321
# of key frames (Serial Search)	236	1012	183
Avg #of comparisons (node fidelity)	184.2	670	146
Avg # of comparisons (edge fidelity)	174.8	532	133

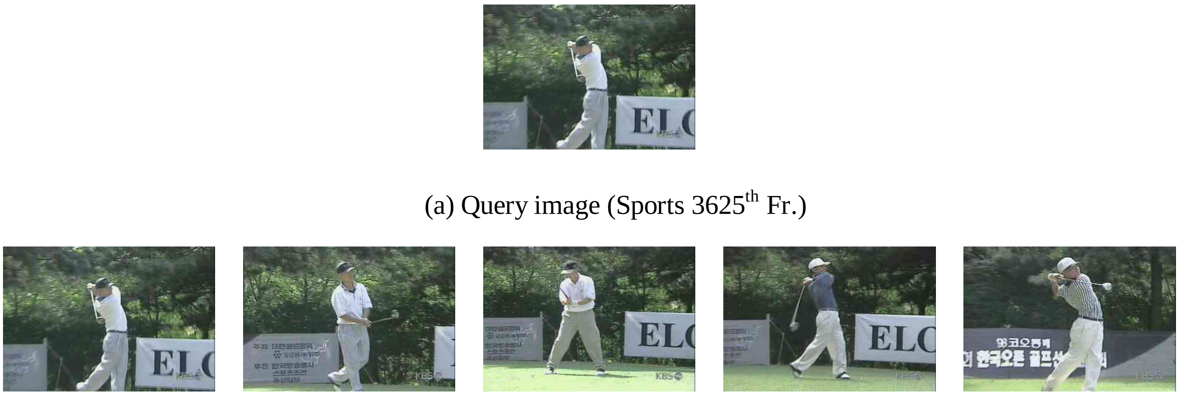
Test video set: Movie clip(True lies), Sports(KBS Golf), Documentary(NHK Documentary)

Fig. 7 shows the performance, when all of the three videos are searched for. The performance of the edge fidelity is experimentally found out to be 12% better than that of the node fidelity. Fig. 8(b) shows a search result for the query image shown in Fig. 8(a).



(a) (b)

Fig. 7. Search performance when searching for all of three videos (a) number of image comparisons in two schemes, (b) improved ratio of search speed of the edge fidelity over the node fidelity.



(a) Query image (Sports 3625th Fr.)

(b) Search result : Sports 3625th Fr.(0.0000), Sports 3155th Fr.(0.1034), Sports 1569th Fr.(0.1299), Sports 15405th Fr.(0.1960), Sports 14046th Fr.(0.2132). Each number in parenthesis indicates the distance between the query image and the retrieved image.

Fig. 8. Retrieval result for a given query image.

5. CONCLUSION

We have proposed a novel scheme allowing for both scalable hierarchical video summary and efficient retrieval by introducing the notion of fidelity. From experiments, we obtained the high performance scalable hierarchical summary, and also found out that the edge fidelity yields 12 % better video search efficiency than the node fidelity, and 30% better than the serial search. In the future, we plan to investigate the performance of the bottom-up clustering used in our current implementation over the top-down clustering.

ACKNOWLEDGEMENT

This work was supported in part by grant No.(98-0102-04-01-3) from the Basic Research Program of the Korea Science & Engineering Foundation.

REFERENCES

1. Hyun Sung Chang, Sanghoon Sull and Sang Uk Lee, "Efficient Video Indexing Scheme for Content-based Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, No. 8, pp.1269-1279, Dec. 1999.
2. Pope A., Kumar R., Sawhney H. and Wan C. Signals, "Video abstraction: summarizing video content for retrieval and visualization," *Systems & Computers, 1998. Conf. Record of the Thirty-Second Asilomar Conf.,* vol. 1, pp. 915 - 919, 1998
3. Macer, P.J. and Thomas, P.J., "Video storyboards: summarizing video sequences for indexing and searching of video databases," *Intelligent Image Databases, IEE Colloquium*, pp. 2/1 -2/5, 1996
4. Uchihashi, S, Foote, J., "Summarizing video using a shot importance measure and a frame-packing algorithm," *Acoustics, Speech, and Signal Processing, 1999. Proceedings*. IEEE International, 1999
5. Jau-Yuen, Cuneyt Taskiran, Edward J. Delp and Charles A. Bouman, "ViBE:A new Paradigm for Video Database Browsing and Search," *Proc. of the IEEE Workshop on Content-Based Access of Image & Video Libraries*, Santa Barbara, CA, June 21, 1998.
6. Minerva M. Yeung and Bede Liu, "Efficient matching and clustering of video shots," *Proc. IEEE International Conf. on Image Processing '95*, vol. 1, pp. 338-341, Oct. 1995.
7. J.Huang, S.R.Kumar, M.Mitra, W.J.Zhu, and R.Zabih, "Image indexing using color correlograms," *Proc. of 16th IEEE Conf. on Computer Vision and Pattern Recognition*, pp.762-768, 1997.
8. Y.Tonomura, A.Akutsu, K.Otsuji, T.Sadakata, "VideoMAP and VideoSpaceIcon : Tools for Anatomizing Video Content," *Proc. ACM INTERCHI '93*, pp.131-141, 1993.
9. A.Nagasaka and Y.Tanaka, "Automatic video indexing and full-video search for object appearances," *Visual Database System II*, 1992.
10. H.Ueda, T.Miyatake, and S.Yoshizawa, "An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System," *Proc. ACM SIGCHI 91, New Orleans, LA*, pp.343-351, 1991.
11. A.M. Ferman and A.M.Tekalp, "Multiscale Content Extraction and Representation for Video Indexing," *Multimedia Storage and Archiving Systems II, Proc. SPIE 3229, Dallas, TX*, pp.23-31, 1997.
12. Y.Taniguchi, A.Akutsu, and Y.Tonomura, "PanoramaExcerpts : Extracting and Packing Panoramas for Video Browsing," *Proc. ACM Multimedia 97, Seattle, WA*, pp.427-436, 1997.
13. Yueting Zhuang, Yong Rui, Thomas S. Huang, and Sharad Mehrotra, "Adaptive key frame extraction using unsupervised clustering," *Proc. IEEE International Conf. On Image Processing*, 1998.
14. Wayne Wolf, "Key frame selection by motion analysis", *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, 1996.
15. Jau-Yuen, Charles A. Bouman and John C. Dalton, "Hierarchical Browsing and Search of Large Databases," *IEEE Trans. on Image Processing*, vol. 9, no. 3, pp. 422-455, 2000.
16. Sanghoon Sull, Jung-Rim Kim, Yunam Kim, "Efficient and effective search and browsing using fidelity," *ISO/IEC JTC1/SC29/SG11 M5101*, Oct. 1999.
17. Sanghoon Sull, Jung-Rim Kim, Yunam Kim, "Improved notion of the fidelity for efficient browsing," *ISO/IEC JTC1/SC29/SG11 M5442*, Dec. 1999.