Proceedings of ITC-CSCC '98, July 13-15, Sokcho, Korea

Set-Theoretic Approach to Video Key Frame Extraction

Hyun-Sung Chang *, **Sanghoon Sull** †, **and Sang Uk Lee**

Signal Processing Lab., School of Electrical Eng., Seoul National University
San 56-1 Shillim-dong, Kwanak-gu, Seoul, 151-742 KOREA
Phone: +82-2-880-8428, Fax: +82-2-880-8220, E-MAIL : hschang@claudia.snu.ac.kr
† Digital Media Lab, School of Electrical Eng., Korea University, Seoul, sull@mpeg.korea.ac.kr

Abstract— **Extracting a small number of key frames which can represent the content of a video shot is important for efficient video browsing and search in video databases. In this paper, we propose an optimal method for extracting key frames based on point set theory. It can find the minimal set of key frames for given degree of accuracy to represent a video shot. A suboptimal Greedy approach is also proposed to speed up the proposed method. Experimental results on a variety of video sequences demonstrate that the proposed method gives about 30% better results in comparison with the existing one, in terms of the R-D performance.**

I. INTRODUCTION

Key frames are a set of images which represent the content of a video shot, where a shot can be defined as a set of the successive frames in video whose content does not change too much. The key frames can be used for video browsing and content-based retrieval and search. For the purpose of retrieval, the similarity of two video shots are often evaluated, based on their key frames. Thus, for large video database, it is desirable to find the minimum number of representative frames which maintain the important content of the video shot.

The simplest existing method for the extraction of key frames is to arbitrarily choose one frame for each shot [1]. More advanced methods [2], [3], [4] have been also proposed by taking into account of the temporal variations within individual shots. These methods attempt to balance the amount of information against the number of key frames. However, most of the approaches are based on heuristics. Moreover, these methods are also constrained by the sequential order of frames, yielding the degradation in the performance.

In this paper, we propose an optimal approach to key frame extraction. The optimality is defined as finding the minimal set of key frames for given degree of accuracy to represent a video shot. We consider the key frame extraction problem as one of choosing the minimal set of points (key frames) from the feature space (video shot), while keeping a predefined distance function less than a given threshold. It is analogous to the vector quantization scheme [5].

The paper is organized as follows. In Section II, we present a new measure of the fidelity of the given key frames for representing a shot. In Section III, we describe a necessary condition for a set of key frames, and propose an optimal algorithm which finds the minimal set of key frames, while satisfying the necessary condition. In Section IV, we show the performance of the proposed algorithm through the experiments. Section V concludes the paper.

II. A NEW MEASURE ON THE FIDELITY OF A SET OF KEY FRAMES

For efficient video browsing and retrieval, selected key frames should be able to represent the content of a shot faithfully. Although there have been several attempts [1], [2], [3], [4] to solve this problem, the proper criterion for the goodness of the selected key frames, which can describe quantitatively the fidelity, has not been proposed yet. In this section, we first present a brief definition of semi-Hausdorff distance. Based on the definition, we propose a new measure for the goodness of the given set of key frames.

Let $(\mathcal{X}, d)$ be a metric space, where d denotes a predefined distance function. For two point sets, $A, B \subset \mathcal{X}$, the semi-Hausdorff distance from A to B, denoted by $d_{SH}(A, B)$, is defined as [6]

$$\begin{aligned} d_{SH}(A,B) &= glb\{\epsilon \mid A \subset U(B,\epsilon)\} \qquad (1) \\ U(B,\epsilon) &= \cup_{x\in B} B_d(x,\epsilon) \\ B_d(x,\epsilon) &= \{y \mid d(x,y) \leq \epsilon\}, \end{aligned}$$

where *glb* represents the *greatest lower bound.*

Let us denote $\{f_1, f_2, f_3, f_4\}$ by S and partition S into two sets, $A = \{f_2, f_3\}$ and $B = \{f_1, f_4\}$ as shown in Fig. 1. Then, from the definition in (1), $d_{SH}(A, B) = 6$.

[*] All correspondence may be addressed to this author.

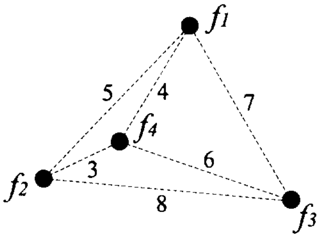

Fig. 1. An example showing the semi-Hausdorff distance. Each number on the line connecting two points indicates the distance between them.

Assuming each frame in a shot corresponds to one feature point, we can apply the definition to the feature space ($\mathcal{I}$,d). Visualizing all the frames in a given shot S are scattered points in $\mathcal{I}$, our goal is to optimally partition S into two groups, R and R^c, which represent a set of key frames and its complement, respectively. Now, we propose $d_{SH}(R^c, R)$ as a measure on the fidelity of selected key frames R, based on the following theorem:

$$d_{SH}(R^c, R) > \epsilon \ \textit{ iff } \ \exists f \in R^c \text{ s.t. } \min_{\hat{f} \in R} d(f, \hat{f}) > \epsilon.$$

Although the theorem, whose proof is omitted here, seems to be simple and trivial, it implies that $d_{SH}(R^c, R)$ can be used for a tight measure to determine how well a given set of key frames represents the entire shot.

Returning to the previous example of Fig 1, assume that $R = \{f_1, f_4\}$ and $R^c = \{f_2, f_3\}$. Then, the fact that $d_{SH}(R^c, R)$ is equal to 6 can be interpreted as follows: The two frames f_1 and f_4 in S can be perfectly represented by the given set of key frames, $R = \{f_1, f_4\}$, which is evident. The remaining two frames f_2 and f_3 should be also represented by R. The f_2 can be approximated by f_4 better than by f_1, and the approximation error is $d(f_4, f_2) = 3$. The f_3 is approximated by f_4 with an error 6, and the error resulting from the approximation of S by R is 6, which is equal to $d_{SH}(R^c, R)$.

Generally $d_{SH}(R^c, R)$ is reduced by utilizing more frames as key frames. For example, if $R = S$, then $d_{SH}(R^c, R) = 0$. Consequently, we are interested in selecting the minimal set of key frames, while maintaining $d_{SH}(\cdot)$ below a predefined threshold ϵ.

III. OUR APPROACH TO KEY FRAME EXTRACTION

In this section, we will describe a necessary condition which a set of key frames should satisfy, and present an optimal method of choosing the minimal set of key frames, while satisfying the condition at the same time.

TABLE I

KEY FRAME SELECTION SCHEMES ALLOWED FOR $d_{SH}(\cdot)$ NOT LARGER THAN 4

R	R^c	$d_{SH}(R^c, R)$	Critical pair
$\{f_3, f_4\}$	$\{f_1, f_2\}$	4	(f_4, f_1)
$\{f_1, f_2, f_3\}$	$\{f_4\}$	3	(f_2, f_4)
$\{f_1, f_3, f_4\}$	$\{f_2\}$	3	(f_4, f_2)
$\{f_2, f_3, f_4\}$	$\{f_1\}$	4	(f_4, f_1)
$\{f_1, f_2, f_3, f_4\}$	ϕ	0	*None*

To reduce the computational complexity of the optimal method, we also present a suboptimal Greedy method.

A. Necessary Condition for a Set of Key Frames

If we let C_i be $\{f_j \in S \mid d(f_i, f_j) \leq \epsilon\}$, a set of key frames R should satisfy the following condition:

$$\cup_{f_i \in R} C_i = S \qquad (2)$$

This condition naturally stems from the definition of semi-Hausdorff distance discussed in section II. If we set ϵ, the supremum of $d_{SH}(\cdot)$, to 4 in Fig. 1, only five out of sixteen possible cases, which are listed in Table I, are allowed because other cases do not satisfy (2). Among this, only the first is optimal, in a sense that the cardinality of R, denoted by $card(R)$, is the smallest, while satisfying the necessary condition. For another example shown in Fig. 2, $F = \{f_3, f_6, f_9, f_{12}\}$ satisfies the condition (2), qualifying to be a set of key frames.

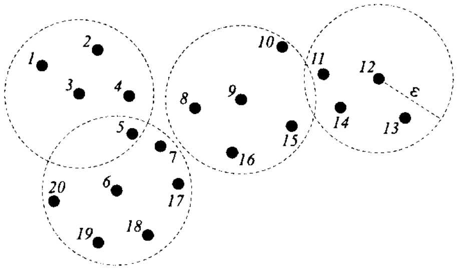

Fig. 2. Twenty frames in feature space $(\mathcal{I}, d)$ corresponding to a shot S. Each number indicates the frame number.

B. An Optimal Approach

Our goal is to find the minimal set of key frames under the constraint that $d_{SH}(\cdot)$ is less than a given threshold. First, we construct a proximity graph as shown in Fig. 3, where each frame within the shot S corresponds to a vertex and two vertices whose distance or cost is less than

ϵ are connected by edges. A vertex is said to cover the neighboring vertices, if the distances to the neighbors are less than ϵ. Then, the minimal set of vertices covering the whole graph becomes the desired solution. The solution can be obtained by Quine-McCluskey algorithm, which is popular in computer logic design, after constructing the covering table as shown in Fig. 4. The details of Quine-McCluskey algorithm can be found in [7]. To reduce the size of the covering table, the techniques, such as row and column dominance, can be used.

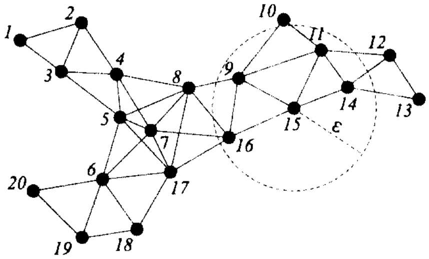

Fig. 3. Proximity graph generated by the parameter ϵ for Fig. 2.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}	f_{20}
C_1	x	x	x																	
C_2	x	x	x	x																
C_3	x	x	x	x	x															
C_4		x	x	x	x		x	x												
C_5			x	x	x	x	x	x									x			
C_6					x	x	x										x	x	x	x
⋮																				
C_{19}						x												x	x	x
C_{20}						x													x	x

Fig. 4. Covering table constructed for Fig. 3

C. A Greedy Approach

Since the finding a minimal set of covering is known to be an NP-complete problem, the use of the proposed optimal method for a video shot containing large number of frames is impractical. To solve this problem, we present a greedy approach whose computational cost is significantly lower with slight degradation in the performance. Let R and C be the set of key frames and the set of frames covered by R, respectively. Let us also define the degree of a vertex f_i as the number of frames in $C_i \cap C^c$.

```
1. $R \leftarrow \phi$,  $C \leftarrow \phi$.
2. Select $f_i \in R^c$ with maximum degree.
   $R \leftarrow R \cup \{f_i\}$,  $C \leftarrow C \cup C_i$.
3. Renew the degrees of vertices, and repeat 2 and 3
   until $C = S$.
```

TABLE II
VARIOUS VIDEO SEQUENCES USED IN EXPERIMENTS

Video sequence	#frames	#shots	min:sec
Commercials	912	16	0:30
Movie Clip	1431	50	1:00
Music Video	6686	163	3:43
News	25974	233	14:26
Sports	27000	181	15:00

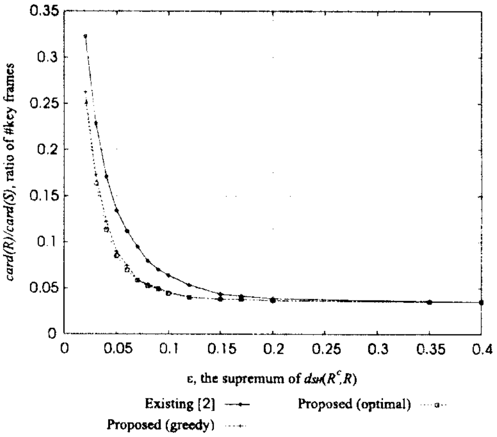

Fig. 5. R-D curve for a movie clip, "True Lies"

Then the obtained R becomes a set of key frames satisfying the necessary condition.

IV. EXPERIMENTAL RESULTS

Experiments have been performed on several MPEG-1 video sequences as listed in Table II. To reduce both temporal and spatial complexities, we use the DC images without decompression [8]. We use the same distance function based on luminance projection vectors (See Appendix I) as in [2]. Note that the optimality of the proposed method is independent of the metric used. To demonstrate the performance, we use a R-D curve where the horizontal axis indicates the maximum of the allowed $d_{SH}(\cdot)$ (i.e. *distortion*), and the vertical axis represents the ratio of the key frames. It is analogous to the R-D curve in the area of source coding. From Fig. 5, we can see that the proposed method yields better performance than the existing one [2]. In fact, the proposed method always achieves the lowest bound.

In Fig. 6, we show an example of video shot to visually demonstrate the superiority of the proposed approach. The conventional methods, which are con-

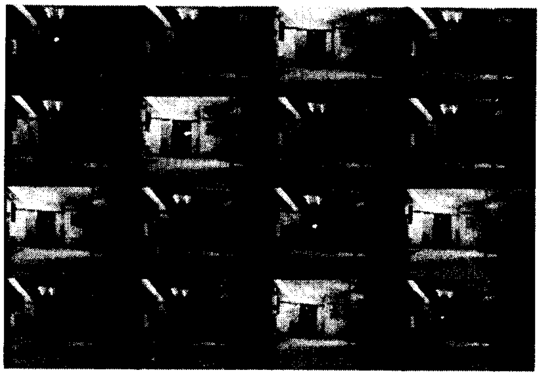

Fig. 6. A shot in "True Lies". From left to right, top to bottom. Each frame is numbered from f_1 to f_{16}.

TABLE III

Results of Key Frame Extraction on 5 Types of Video Sequences. The Greedy Algorithm is Used.

Video sequence	ϵ	$card(R)/card(S)(\%)$	
		Existing [2]	Proposed
Commercials	0.03	4.39	3.07
Movie Clip	0.04	16.98	12.16
Music Video	0.04	20.43	14.94
News	0.04	3.82	2.50
Sports	0.04	6.73	4.44

strained by the sequential order of frames, select the key frames whenever the luminance changes. That is, $R = \{f_1, f_3, f_4, f_6, f_7, f_9, f_{10}, f_{12}, f_{13}, f_{15}, f_{16}\}$. On the other hand, the proposed method selects only two frames(f_3 and f_4); One in light mood and the other in dark mood. It is found that the proposed method works well for many other cases than that due to luminance effects. Table III shows the experimental results for 5 different types of video sequences. We see that the proposed method is about 30% better than the existing one at $\epsilon = 0.03 \sim 0.04$(typical value) in R-D performance.

V. CONCLUSION

For efficient video browsing and content-based retrieval, based on a set-theoretic approach, we proposed an optimal method for selecting a minimal set of key frames for a given threshold. First, we have presented a new quantitative measure on the goodness of a set of the key frames selected from a given video shot. Second, we have described an optimal method based on the measure. Then, we have presented a suboptimal Greedy approach to speed up the proposed method. The experimental results on a variety of video sequences showed that the proposed method provides a superior R-D performance, compared to the previous approaches [2]. The proposed method can be used for fast video search with considerable accuracy. It will be also useful to clustering shots in video structuring process.

References

[1] F.Arman, R.Depommier, A.Hsu and M-Y.Chiu, "Content-based browsing of video sequences", in *Proc. ACM Multimedia Conf.*, pp.97-103, Aug. 1994.

[2] M.M.Yeung and B.Liu, "Efficient matching and clustering of video shots", in *Proc. Int. Conf. Image Processing*, vol.1, pp.338-341, Oct. 1995.

[3] R.L.Lagendijk *et al.,* "Visual search in a SMASH system", in *Proc. Int. Conf. Image Processing*, vol.3, pp.671-674, Sep. 1996.

[4] H.Aoki, S.Shimotsuji, and O.Hori, "A shot classification method of selecting effective key-frames for video browsing", in *Proc. ACM Multimedia Conf.*, pp.1-10, Nov. 1996.

[5] A.Gersho and R.M.Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[6] J.R.Munkres, *Topology: A First Course,* Prentice Hall, 1975.

[7] J.P.Hayes, *Introduction to Digital Logic Design,* Addison Wesley, 1993.

[8] B.Yeo and B.Liu, "Rapid scene analysis on compressed video", *IEEE Trans. Circuits Syst. Video Technol.*, vol.5, pp.533-544, Dec. 1995.

Appendix

I. DISTANCE FUNCTION USING LUMINANCE PROJECTION VECTORS

For a given image $f(j,k), j = 1, 2, \cdots, J$ and $k = 1, 2, \cdots, K$, the luminance projection vectors for the nth row and mth column, denoted by l_n^r and l_m^c respectively, are defined as

$$\begin{aligned} l_n^r &= \sum_{j=1}^{J} Lum\{f(j,n)\} \\ l_m^c &= \sum_{k=1}^{K} Lum\{f(m,k)\} \end{aligned}$$

Then the distance function $d_{lp}(\cdot)$, normalized to [0,1], is defined as follows [2]:

$$\begin{aligned} d_{lp}(f_i, f_j) = \frac{1}{255(J+K)}(\frac{1}{J}\sum_{n=1}^{K} \mid l^r(f_i)_n - l^r(f_j)_n \mid \\ + \frac{1}{K}\sum_{m=1}^{J} \mid l^c(f_i)_m - l^c(f_j)_m \mid) \end{aligned}$$

The performance using the luminance projection vectors is reported to be comparable to that of using correlation between two frames.