Equivalent Key Frames Selection Based on Iso-Content Distance and Iso-Distortion Principles

Costas Panagiotakis, Anastasios Doulamis and Georgios Tziritas
Multimedia Informatics Laboratory of Computer Science Department, University of Crete
Heraklion, P.O. Box 2208, 71409, Greece
phone: + (30) 2810 393517, fax: + (30) 2810 393501
e-mail: {cpanag, adoulam, tziritas}@csd.uoc.gr

***Abstract*— We present a key frames selection algorithm, which is very flexible on any changes of content descriptors, based on Iso-Content Distance and Iso-Distortion principles. In both of the cases, the equality principle provides to the selected key frames the property to be equivalent on content video summarization. The estimated key frames properties and the experimental results indicate the good performance of the proposed schemata.**

I. INTRODUCTION

The most of key frames selection techniques assume that the video has been segmented into shots and then extract within each shot a small number of representative key frames. A shot can be defined as a sequence of frames that are (or appears to be) continuously captured from the same camera. Key frames can be defined as a subset of a video sequence that can represent the video visual content as close as possible, with a limiting number of frames [1].

Before applying a video summarization algorithm, appropriate visual features are extracted from each video file [2]. Visual content descriptors like color-texture descriptors, color-edge histograms, motion vectors have been used in key frames selection methods [3].

Key frames selection approaches can be classified into: cluster-based methods, energy minimization-based methods and sequential methods. The clustering techniques [4] take all the frames of a shot together and classify them according to their content similarity. Then, the key frames are determined as the representative frame of a cluster. The disadvantage of these approaches is that they ignore the temporal information of a video sequence. Thus, the selected key frames can not be used in video similarity and indexing based applications. The energy minimization based methods [5] extract the key frames by solving an energy minimization problem. These methods are generally computational expensive using iterative techniques to perform minimization. The sequential methods [6] consider a new key frame when the content difference from the previous key frame exceed a a predefined threshold that is determined by the user. Three approaches for video summarization has been proposed in [7]. All approaches minimize a cross-correlation criterion so that the most uncorrelated frames as represented in the feature domain to be extracted as the most representative. Since the complexity of an exhaustive search is very high especially for long shots and high number of key-frames, a logarithmic, stochastic logarithmic and a genetic approach has been proposed in [7] to improve the search efficiency. Extension of [7] to stereoscopic data has been proposed in [8].

All the above mentioned approaches address the video summarization problem focusing either on a restricted video content, ignoring temporal variation, minimizing metric criteria on feature domain, or applying simple clustering-based techniques. On the contrary, in this paper, video summarization is performed by the use of an innovative computational geometry algorithm, which equally partitions the *content curve* of a video sequence resulting in key frames that are *equivalent* in the content domain under any type of video content description ([9]–[11]). In this paper, we propose two general principles based on this algorithm that can be used in key frames selection. Applying the equipartition algorithm directly on content description, we get equidistant key frames in the sense of video content, named Iso-Content Distance principle. Alternatively, under Iso-Content Distortion principle the selected key frames minimizes a global distortion criterion providing at the same time equal distortions per key frame cluster. The main contribution of this work, is to address the problem of video summarization from different views, the proposed Iso-Distance and Iso-Distortion principles, that take into account the equivalent property of the key frames under any type of content description.

The rest of the paper is organized as follows: Section II gives the problem formulation describing the proposed principles. Section III presents the key frames selection algorithm. The visual content descriptors are presented in Section IV. The experimental results are given in Section V. Finally, conclusions are provided in Section VI.

II. PROBLEM FORMULATION

Let us assume a video shot of N frames duration and that for each frame of the shot, we have extracted several descriptors and included them in a vector denoted as $\mathbf{p_i}$, where index i corresponds to the ith frame shot. Let us denote as P a set which includes all vectors $\mathbf{p_i}$ for $i = 1, 2, \cdots, N$, that is $P = \{\mathbf{p_1}, \mathbf{p_2}, \cdots, \mathbf{p_N}\}$. Vectors $\mathbf{p_i}$ are assumed to be $\in R^n$, i.e., n descriptors are extracted to represent its frame content.

According to EP problem, we have to use as input a continuous time descriptor curve $C(t)$, where t denotes the time variable, instead of the set P. Therefore, $C(t)$ can be derived

by the linear interpolation of the successive frames descriptors in n dimensional space. To simplify the mathematical formulations and without loosing generality, we have normalized the time variable t, $t \in [0,1]$, so that $0, 1$ correspond to first, last frames, respectively. Thus, we assume that $C(t)$ starts on $A = C(0) = \mathbf{p_1}$ and ends on $B = C(1) = \mathbf{p_N}$. In the next sections, we are going to keep using the continuous normalized time space $[0,1]$ instead of the discrete frames' time space $\{1, 2, \cdots, N\}$.

The EP problem is defined under a predefined smooth semimetric function like Euclidean distance. By our analysis [9]–[11], the equipartition problem (EP) admits always a solution under any semimetric function. Therefore we have to use as $g(x,y)$ a semimetric function in order to get at least one solution. Let $g(x,y)$, where $x, y \in [0,1]$ denote normalized time variables, be the used smooth semimetric function between two curve points $C(x)$, $C(y)$.

Under the key frames selection problem, the key frames are selected to summarize the video content. Let M be the number of the selected key frames and $t'_i \in [0,1], i \in \{1, \cdots, M\}$ be the selected key frames under the normalized time space. The proposed method selects the first and the last key frame to be the first and the last frame of the shot sequence, respectively ($t'_0 = 0, t'_1 = 1$). Therefore, the goal of the proposed method is to compute $M-2$ key frames K, $t'_i, i \in \{2, \cdots, M-1\}$, under the constraint that are equidistant in the sense of the used semimetric function $g(x,y)$,$g(t'_{i-1}, t'_i) = g(t'_i, t'_{i+1}),\ i \in \{2, \cdots, M-1\}$, with $t'_1 = 0$ and $t'_M = 1$. This means that the distance between each successive pair of key frames will be equal. The length r of each equal chord is given by the following equation:

$$r = g(0, t'_2) = g(t'_2, t'_3) = \cdots = g(t'_{M-1}, 1) \qquad (1)$$

Therefore, the set of key frames $K = \{t'_1, t'_2, \cdots, t'_M\},\ t'_i < t'_{i+1},\ C(t'_1) = \mathbf{p_1}, C(t'_M) = \mathbf{p_N}$ are selected under any predefined content description having equivalent property on video content descriptors. We propose two general principles that can be used in key frames selection: the Iso-Content Distance and the Iso-Content Distortion principles.

A. Iso-Content Distance Principle

Under Iso-Content Distance principle, the content distances between two successive key frames should be equal, so the selected key frames are equidistant in content. Thus under the definition of Section II, we have to compute $M-2$ sequential key frames $t'_i, i \in \{2, \cdots, M-1\},\ t'_1 = 0,\ t'_M = 1$ under the constraint: $r = d(t'_{i-1}, t'_i) = d(t'_i, t'_{i+1}),\ i \in \{2, \cdots, M-1\}$, where $d(x,y),\ x, y \in [0,1]$ denotes the semimetric distance function. Several distances $d(x,y)$ can be used, like Euclidean, Manhattan, χ^2, depending on content descriptors' space. If there are several solutions, the one with the maximum chord length r is selected since this solution is the best approximation of the content curve.

B. Iso-Content Distortion Principle

Under Iso-Content Distortion principle, the distortions between two pairs of key frames, $\bar{d}(t'_i, t'_{i+1})$= $\bar{d}(t'_j, t'_{j+1})$ should be equal. We consider the following definition for distortion: Let $t'_i,\ t'_{i+1}$ be two successive key frames, then the distortion $\bar{d}(t'_i, t'_{i+1})$ is defined as the sum of minimum content distances of the frames $t_j,\ t'_i \le t_j \le t'_{i+1}$ and the two key frames $t'_i,\ t'_{i+1}$,

$$\bar{d}(t'_i, t'_{i+1}) = \sum_{j=t'_i}^{t'_{i+1}} \min(d(t'_i, t_j), d(t_j, t'_{i+1})). \qquad (2)$$

This is a similar definition with the definition of distortion used by Lee and Kim [5],$\bar{d}(K,B) = \sum_{i=1}^{M} \int_{b_i}^{b_{i+1}} d(t, t'_i)dt$. Breakpoints have not been used ($B = \{b_0, \cdots, b_M\}$), since their meaning is included in the two successive key frames combination in Eq. (2). If we define the total distortion as the maximum of the corresponding distortions $\max_{i \in \{1,2,\cdots,M-1\}} \bar{d}(t'_i, t'_{i+1})$, then almost optimal solutions are achieved using the proposed schema. If there are several solutions, the one with the minimum distortion is selected.

III. Key Frames Selection Algorithm

The straightforward implementation of the EP method provides directly the M key frames. The EP algorithm computes for a specific M, the M key frames K under the semimetric $g(x,y)$ function. The number of key frames M can be given by the user or can be estimated automatically by exploiting the variation of feature vector trajectory in time [12].

The input of the method is the number of key frames M. In addition, it needs the values of symmetric matrix $g(t_k, t_l), k, l \in \{1, 2, \cdots, N\}$. This algorithm is described in [9], [11] computing all solutions in $O(M \cdot N^2)$ steps. A brief description of EP algorithm is given next. It is an iterative method. Thus, when it is executed for M segments, it uses the precomputed results for $M-1$ segments. In each iteration step l, the algorithm computes the zero level curves L_l by the L_{l-1}. These curves points belong on the same level of $g(x,y)$ and the key frames are inductively computed (from L_l to L_{l-1}) on these curves (see Figs. 4(a) and 4(g)). By our analysis [9]–[11], the equipartition problem (EP) admits always a solution. The EP problem can have more than one solutions depending on curve shape, distance metric, and the value of M.

IV. Visual Content Description

The proposed method can be executed under any choice or combination of audio/visual content descriptors. However, the selected key frames are related with the used content description, so we have to choose appropriate descriptors. On this framework, we propose to use the MPEG-7 visual descriptors [3] like the Color Layout Descriptor (CLD), a low cost and compact descriptor, which suffices to describe smoothly the changes in visual content of a shot. We used the following function D to measure the content distance of two CLDs, $\{DY, DCb, DCr\}$ and $\{DY', DCb', DCr'\}$, $D = \sqrt{\sum_i (DY_i - DY'_i)^2} + \sqrt{\sum_i (DCb_i - DCb'_i)^2} + \sqrt{\sum_i (DCr_i - DCr'_i)^2}$, where, (DY, DCb, DCr) represent the ith DCT coefficients of the respective color components. The function D is a semimetric distance.

(a) #0 (b) #97 (c) #132 (d) #299

Fig. 1. The proposed key frames are $\{0, 97, 299\}$ and $\{0, 132, 299\}$ in coast shot under Iso-Content Distance and Iso-Distortion principles, respectively.

V. Experimental Results

In this section, the experimental results of the proposed algorithm and comparisons to other algorithms are presented. The method has been implemented using C and Matlab.

A. Evaluation of the Proposed Schemata

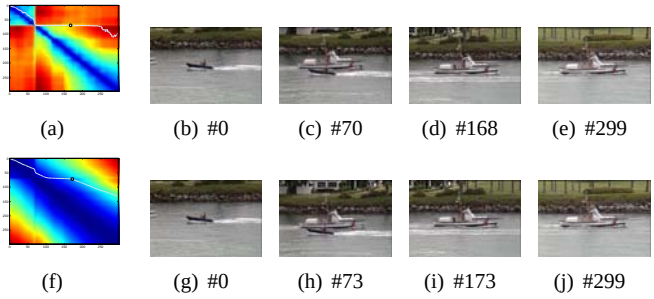
We have tested the proposed algorithm on a data set containing more than 250 video sequences. The most of them are athletics videos like pole vault, high jump, triple jump, long jump, running and hurdling. Moreover, we have used the widely known as MPEG test sequences like coast sequence, the table tennis sequence, hall monitor sequence, etc.. Figs. 5 and 6, show the sequences that we used in the article.

A typical processing time for the execution of the proposed EP algorithm, when the shot contains 300 images (e.g. coast MPEG sequence) and $M = 10$, is between 4 to 5 seconds depending on the used principle. Figs. (4(a), 4(g)) and (2(a), 2(f)) show the surfaces $g(x, y)$ in pole vault and coast sequences, respectively, under the proposed principles. The deep blue colors correspond to close to zero values. This is the reason of the deep blue diagonals, since it holds that $g(x, x) = 0$. The deep red colors correspond to the highest values of $g(x, y)$. The estimated solution is projected on $g(x, y)$ with cycles. The L_l curves are projected on $g(x, y)$, with gray colors, at both sides of diagonal $x = y$. If more than one solutions are appeared, the selected solution points are drawn with large cycles.

Figs. 2 and 4 illustrate the results of the two proposed schemata in pole vault and coast sequences, respectively. The number of key frames has been automatically estimated using the criterion of [12]. We have observed that, under Iso-Content Distortion principle the better representative frames of their cluster are selected. Moreover, the selected key frames under Iso-Content Distance principle don't take the duration between the selected key frames into account, while the Iso-Content Distortion combines the duration with the content variation (Fig. 1). Moreover, we have tested the two proposed schemata for low number of key frames (see Fig. 1), indicating the robustness of the method.

B. Comparison to Other Algorithms

The proposed scheme has been compared with two approaches presented in the literature. The first exploits a minimization of a cross correlation criterion [7], so that the most uncorrelated frames are extracted as key ones. A logarithmic search approach is adopted as in [7] to estimate the key frames. The second technique formulate the summarization problem as an interpolation problem. Fig. 3 illustrates the performance of both methods along with the proposed one for the pole vault sequence. It is observed that the proposed approach represents the content of the sequence more efficiently rather than the compared works. In all case, the same number of key-frames has been extracted obtained using the criterion of [12].



(a) (b) #0 (c) #70 (d) #168 (e) #299

(f) (g) #0 (h) #73 (i) #173 (j) #299

Fig. 2. Results of the two proposed schemas in coast shot using four key frames. The estimated solution and the L_l curves are projected on $g(x, y)$ under **(a)** Iso-Content Distance, **(f)** Iso-Content Distortion principle. **(b),** $\cdots$, **(e)** The selected key frames under Iso-Content Distance principle. **(g),** $\cdots$, **(j)** The selected key frames under Iso-Content Distortion principle.

VI. Conclusions

In this paper, two key frames selection schemata are described based on equipartition problem. The first one uses Iso-Content Distance principle, the key frames are equidistant in video content. Under Iso-Content Distortion principle, the frames clusters derived by the key frames are equal-sized. Thus, the selected key frames have different properties according to the used principle. However, in any case, the key frames are equivalent on content video summarization. Each key frame has the same significance under the used principle. In this work, we have used the Color Layout Descriptor of MPEG-7 visual descriptors.

An extension of the proposed methodology may include automatic computation of key frame number, the using of more audio/visual descriptors and principles.

Acknowledgment

This research was partially supported by the Greek PENED 2003 project.

References

[1] M. Yeung and B.-L. Yeo, "Video visualization for compact presentation and fast browsing ofpictorial content," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 7, no. 5, pp. 771 – 785, 1997.

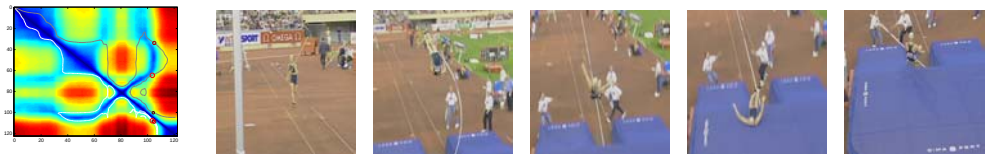
[2] Y.-P. Tan, S. R. Kulkarni, and P. J. Ramadge, "A Framework For Measuring Video Similarity And Its Application To Video Query By Example," 1999.

[3] B. Manjunath, J. Ohm, V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. On Circuits And Systems For Video Tech.*, vol. 11, no. 6, pp. 703–715, 2001.

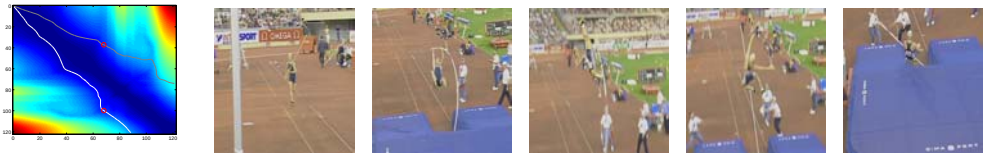
[4] A. Girgensohn and J. S. Boreczky, "Time-constrained keyframe selection technique," *Multimedia Tools and Applications*, vol. 11, no. 3, pp. 347–358, 2000.

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

[6] J. Vermaak, P. Perez, and M. Gangnet, "Rapid summarization and browsing of video sequences," in *British Machine Vision Conf.*, 2002.

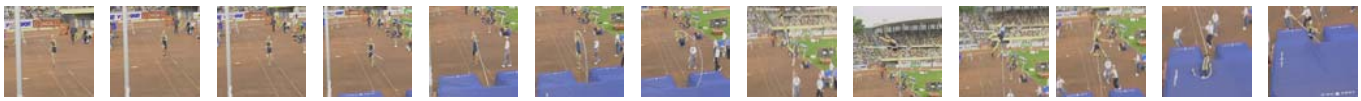


(a) (b) #1 (c) #65 (d) #104 (e) #109 (f) #123



(g) (h) #1 (i) #53 (j) #69 (k) #101 (l) #123

Fig. 4. Results of the two proposed schemas in pole vault shot using five key frames. The estimated solution and the l_t curves are projected on $g(x, y)$ under **(a)** Iso-Content Distance, **(g)** Iso-Content Distortion principle. **(b),** $\cdots$**, (f)** The selected key frames under Iso-Content Distance principle. **(h),** $\cdots$**, (l)** The selected key frames under Iso-Content Distortion principle.



(a) #0 (b) #10 (c) #20 (d) #30 (e) #40 (f) #50 (g) #60 (h) #70 (i) #80 (j) #90 (k) #100 (l) #110 (m) #123

Fig. 5. The pole vault sequence which contains 123 frames.

(a) #0 (b) #15 (c) #30 (d) #45 (e) #60 (f) #75 (g) #90 (h) #105 (i) #120 (j) #135 (k) #150

(l) #165 (m) #180 (n) #195 (o) #210 (p) #225 (q) #240 (r) #255 (s) #270 (t) #285 (u) #299

Fig. 6. The coast sequence which contains 300 frames.

[7] N. Doulamis, A. Doulamis, Y. Avrithis, K. Ntalianis, and S. Kollias, "A stochastic framework for optimal key frame extraction from mpeg video databases," *Journal of Computer Vision and Image Understanding,*, vol. 75, no. 4, pp. 3–24, 1999.

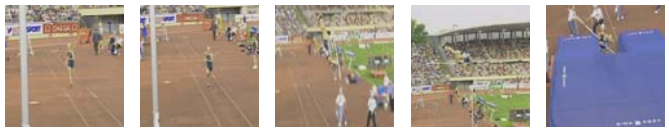
[8] ——, "Efficient summarization of stereoscopic video sequences," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 10, no. 4, pp. 501–517, 2000.

[9] C. Panagiotakis, G. Georgakopoulos, and G. Tziritas, "The curve equipartition problem," *submitted to Computational Geometry*, 2005. [Online]. Available: http://www.csd.uoc.gr/~cpanag/papers/EP.pdf

[10] ——, "On the curve equipartition problem: a brief exposition of basic issues," in *European Workshop on Computational Geometry*, 2006.

[11] C. Panagiotakis and G. Tziritas, "Any dimension polygonal approximation based on equal errors principle," *Pattern Recogn. Lett.*, vol. 28, no. 5, pp. 582–591, 2007.

[12] A. D. Doulamis, N. Doulamis, and S. Kollias, "Non-sequential video content representation using temporal variation of feature vectors," *IEEE Trans. on Consumer Electronics*, vol. 46, pp. 758–768, 2000.



(a) #1 (b) #21 (c) #69 (d) #82 (e) #123

(f) #34 (g) #49 (h) #67 (i) #72 (j) #90

Fig. 3. Results of interpolation and logarithmic method described in [7] in pole vault shot using five key frames. **(a),** $\cdots$**, (e)** The selected key frames under interpolation method. **(f),** $\cdots$**, (j)** The selected key frames under logarithmic method.