MINMAX Video Summarization under Equality Principle

Costas Panagiotakis, Ilias Grinias 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, grinias, tziritas}@csd.uoc.gr

***Abstract*— In this paper we present a video summarization scheme. First, shot detection is performed and then we extract the key frames under the equality principle. We propose a key frames selection algorithm (Iso-Content MINMAX), which is very flexible on any changes of content descriptors, based on MINMAX optimization formulation. The equality principle provides to the selected key frames the useful property to be equivalent on content video summarization.**

I. Introduction

The key frame extraction techniques assume that the video file has been segmented into shots and then they extract within each shot a small number of representative frames (key frames). A shot can be defined as a sequence of frames that are or appear to be continuously captured from the same camera. Ideally, a shot can encompass pans, tilts, zooms or any other camera effects [1]. Thus, shot cut detection algorithms can be considered as the first step towards a video abstraction scheme. A number of shot detection methods have been proposed in the literature using as content description, for example, the difference between the color histograms [2], the Discrete Cosine Transform (DCT) or optical-flow computation [1].

Key frames selection approaches can be classified into cluster-based methods, energy minimization-based methods and sequential methods. The clustering techniques [3] take all the frames of a shot together and classify them according to their content similarity. Then, key frames are determined as the representative frames of a cluster. The disadvantage of these approaches is that they ignore the temporal information of a video sequence. The energy minimization based methods [4] extract the key frames by solving an energy minimization problem. These methods are generally computational expensive, since they use iterative techniques to perform minimization. The sequential methods [5] 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 have been proposed in [6]. All approaches minimize a cross-correlation criterion so that the most uncorrelated frames in feature content domain are considered as the most appropriate key frames.

Recently, dynamic programming techniques have been proposed in the literature, such as the MINMAX approach of [7] to extract the key frames of a video sequence. In this work, the problem is solved optimally in $O(N^2 \cdot K_{max})$, where K_{max} is related to the distortion rate of the solution. In [8], a generic user attention model is proposed and then applied for summarizing video visual data. The model initially models the human's attention through multiple sensory perceptions, i.e., visual and aural stimuli, and then summarizes the data based on this model.

All the above mentioned approaches address the video summarization problem by focusing either on a restricted video content ignoring temporal variation and 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]–[12]), the curve EquiPartition problem (EP). Under equality principle, we MINimize the MAXimum frame distortion (MINMAX) [7] per shot, which is found to be a good metric that matches the subjective perception of the distortion [7]. The main contribution of this work, is that we take into account the equivalent property of the key frames under the MINMAX optimization formulation.

The rest of the paper is organized as follows: Section II describes the shot detection algorithm. Section III gives the problem formulation and the reduction to EP. Section IV presents the key frames selection algorithm. The visual content descriptors are presented in Section V. The experimental results are given in Section VI. Finally, conclusions are provided in Section VII.

II. Shot Detection

Video shot detection refers to the segmentation of video in segments of fairly different visual content. Shot boundaries are defined on the abrupt or gradual transitions among successive video frames. In this work, only the abrupt video content changes among successive video frames f_{t-1} and f_t are detected, while the not detected gradual shot transitions are handled by the keyframes extraction algorithm. Color histograms have been used before for shot detection [2] and it is well known that they are robust against scene activities such as camera and objects motion. Hence, in order to determine an abrupt change, the color histograms of the two frames h_{t-1} and h_t respectively are computed in *YCbCr* color space and

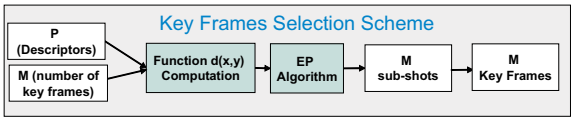


Fig. 1. Scheme of proposed EP algorithm.

a new shot beginning timestamp at t is detected using the χ^2 statistical metric, if

$$\chi^2(h_{t-1}, h_t) = \sum_i \frac{(h_{t-1}[i] - h_t[i])^2}{h_{t-1}[i] + h_t[i]} > \mathcal{T}$$

where $\mathcal{T}$ is a predefined percent of image pixels.

III. Problem formulation and the Reduction to EP

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 the 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.

Let $K = \{t'_1, t'_2, \cdots, t'_M\}$ be a set of key frames and $d(t_i, t_j)$ be the distortion between two frames t_i and t_j. $d(t_i, t_j)$ can be estimated using the descriptor's curve (see Section V). Let $B = \{b'_0, b'_1, \cdots, b'_M\}$, with $b'_0 = 1$, $b'_M = N$ and $b'_{i-1} \leq t'_i \leq b'_i$, $\forall i \in \{1, \cdots, M\}$ be the boundary frames of the key frames. Each sequential couple b'_{i-1}, b'_i defines a sub-shot. The reconstructed video sequence $V'_K = \{u'_1, u'_2, \cdots, u'_N\}$ is obtained from the set of key frames K and the set of their boundaries B, by substituting missing frames with the corresponding key frame of the set K, so that:

$$u'_k = t'_i \Leftrightarrow b'_{i-1} \leq k \leq b'_i \quad (1)$$

In [7], the reconstructed video sequence is obtained from the set of key frames K, by substituting the missing frames with the most recent frame that belongs to set K. According to MINMAX criterion, K is selected in order to minimize the maximum frame distortion between the original sequence and its reconstruction (using the set of key frames).

$$D(K) = max_{k \in \{1, \cdots, N\}} d(t_k, u'_k) \quad (2)$$

A. The EP problem

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 and 1 correspond to first and 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 $\{1, 2, \cdots, N\}$. Thus, $d(x, y)$ will denote the distortion between two curve points $C(x)$, $C(y)$.

The EP problem is defined under a predefined smooth semimetric function[1] like Euclidean distance. $d(x, y)$ is such a semimetric function. The goal of EP is to compute $M - 2$ sequential curve points $x'_i \in [0, 1], i \in \{2, \cdots, M-1\}$, $x'_0 = 0$, $x'_1 = 1$ so that under the constraint that are equidistant in the sense of the used semimetric function $d(x, y)$, $d(x'_{i-1}, x'_i) = d(x'_i, x'_{i+1})$, $i \in \{2, \cdots, M-1\}$, with $x'_1 = 0$ and $x'_M = 1$. This means that the distance between each successive pair of points will be equal. The length r of each equal chord is given by the following equation:

$$r = d(0, x'_2) = d(x'_2, x'_3) = \cdots = d(x'_{M-1}, 1) \quad (3)$$

Figure III illustrates the proposed EP algorithm scheme. In the next section, we describe how the MINMAX is reduced to EP.

B. The reduction of MINMAX to EP

Under EP principle, the "distortions" between two pairs of points, $d(x'_i, x'_{i+1})$= $d(x'_j, x'_{j+1})$ should be equal. The MINMAX problem can be reduced to EP, if we apply it on the set of boundaries B, so that $d(b'_i, b'_{i+1})$= $d(b'_j, b'_{j+1})$. Next, the corresponding set of key frames $K = \{t'_1, t'_2, \cdots, t'_M\}$ can be defined by the set of boundaries B under the minimization of the maximum frame distortion in the corresponding subsequence (sub-shot).

$$t'_i = argmin_{t_k \in \{b'_{i-1}, \cdots, b'_i\}} D(t_k) \quad (4)$$

In many cases almost optimal solutions are obtained. It holds that if the given distortion surface $d(x, y)$ is smooth then the proposed solution with great probability will be the optimal [11]. The video content smoothly changes over a shot time, so the produced $d(x, y)$ will be smooth under the assumption of robustness of the adopted content descriptors. Moreover, the equal distortions per sub-shot mean that the key frames are equivalent in video content summarization.

IV. The Iso-Content MINMAX Key Frames Selection Algorithm

The straightforward implementation of the EP method provides directly $M+1$ boundary points and M key frames. The number of key frames M can be given by the user or can be estimated automatically by terminating the EP algorithm when the estimated distortion exceeds a predefined error, similar with the the problem of minimum number of segments ($min - \#$) [11]. Both of the cases are solved in $O(M \cdot N^2)$ steps thanks to the property of the method that it solves the problem for values less than M without additional cost.

The input of Iso-Content MINMAX method is the number of key frames M. In addition, it needs the values of symmetric matrix $d(t_k, t_l), k, l \in \{1, 2, \cdots, N\}$ of distortions. This algorithm is described in [9], [11]. First all the sub-shots solutions (set B) are extracted and then, we can estimate

[1]A semimetric function is defined as the one which satisfies positive definiteness and symmetry (like a metric function), but it is not required to satisfy the triangular inequality property.

the key frames from their boundaries B using the symmetric matrix $R(b'_{i-1}, b'_i) = argmin_{k \in \{b'_{i-1}, \cdots, b'_i\}} D(t_k)$. 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 curve points belong to the same level of $d(x, y)$ and the key frames' boundaries are inductively computed (from L_l to L_{l-1}) on these curves (Fig. 2). By our analysis [9]–[11], the equipartition problem (EP) always admits a solution.

V. 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 use the MPEG-7 visual descriptors [13] like the Color Layout Descriptor (CLD), a low cost and compact descriptor, which suffices to describe smoothly the changes in visual content (mainly color and motion variations) of a shot. We used the following semimetric 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.

VI. Experimental Results

In this section, the experimental results of the proposed algorithm together with comparisons to other algorithms are presented. We have tested the proposed algorithm on a data set containing more than 500 video sequences. Some of them are athletics videos like pole vault, high jump, triple jump, long jump, running and hurdling. Moreover, we have used a big dataset of television news [2]. Figs. 5 (from athletics) and 6 (from television news), show the sequences that are used in the article.

Fig. 2 shows the surfaces $d(x, y)$ in pole vault and snowing sequences, respectively. The deep blue colors correspond too close to zero values. This is the reason of the deep blue diagonals, since it holds that $d(x, x) = 0$. The deep red colors correspond to the highest values of $d(x, y)$. The estimated solution (key frames boundaries) is projected on $d(x, y)$ with cycles. The L_l curves are projected on $d(x, y)$, with gray colors, at both sides of diagonal $x = y$. Figs. 3 and 4 illustrate the results of the proposed algorithm in pole vault and snowing sequences, respectively.

The proposed scheme has been compared with the Iso-Content Distance and Iso-Content Distortion principles proposed in [12] which has been found that outperform the methods of minimization of cross correlation criterion [6]. Under Iso-Content Distance principle, the key frames are equidistant in video content. Under Iso-Content Distortion principle, the frame clusters derived by the key frames are equal-sized under a similar criterion proposed by Lee and Kim [4]. Figs. 3 and 4 illustrate the performance of both methods compared to the proposed one for the pole vault and snowing sequence. To derive a fair comparison, the same number of frames are used. We have observed that, under Iso-Content MINMAX principle the better representative frames of their sub-shots are selected, yielding more key frames during the highest part of the jump, which is the best solution in the sense of human perception. Moreover it gives more key frames during the end of the snowing video, where the visual content variation is maximized due to a very fast camera motion. Under Iso-Content Distortion principle high representative frames are selected taking into account the time duration between the key frames yielding more equal-time-distributed key frames. However, the using of time duration make the method less flexible to provide more key frames when the content actually varies (see Figs. 3 and 4). The selected key frames under Iso-Content Distance and Iso-Content MINMAX principles don't take the duration between the selected key frames into account. Under Iso-Content Distance the variation of the intrinsic content curve between the key frames is ignored, which can cause the ignorance of an important visual sub-curve (e.g. highest part of the jump in pole vault video). Moreover, under the Iso-Content MINMAX principle it is not required to add in the set of key frames the first and the last frame of the sequence, which is required under Iso-Content Distance and Iso-Distortion principles.

[2]The national Greek television (ET-3) supports us with a 50 hour video-dataset of news.

VII. Conclusions

In this paper, we propose a video summarization scheme based on a shot detection algorithm and a MINMAX key frames selection technique which is based on equipartition principle. For each shot of the video sequence, the frame clusters derived by the key frames boundaries are equal-sized. In each sub-shot, the most representative frame is selected according to the minimization of the maximum frame distortion of the sub-shot. Therefore, the key frames are equivalent on content video summarization. By our experimental results, it is observed that the proposed approach represents the content of the sequence more efficiently rather than the compared works due to the MINMAX criterion. An extension of the proposed methodology may include the use of more audio/visual descriptors and a weighting combination of distances under equality principle.

Acknowledgment

This research was partially supported by the Greek PENED 2003 and TV++ projects. The authors would like to thank the researchers of LIS (Image and Signal processing Lab) at Grenoble, Emmanuel Ramasso, Michèle Rombaut and Denis Pellerin for the data exchange.

References

[1] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods." *IEEE Trans. Circuits Syst. Video Techn.*, vol. 10, no. 1, pp. 1–13, 2000.

(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.

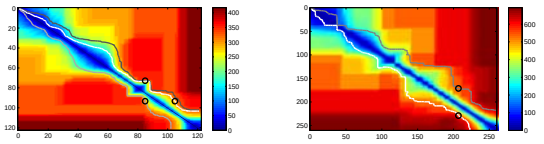
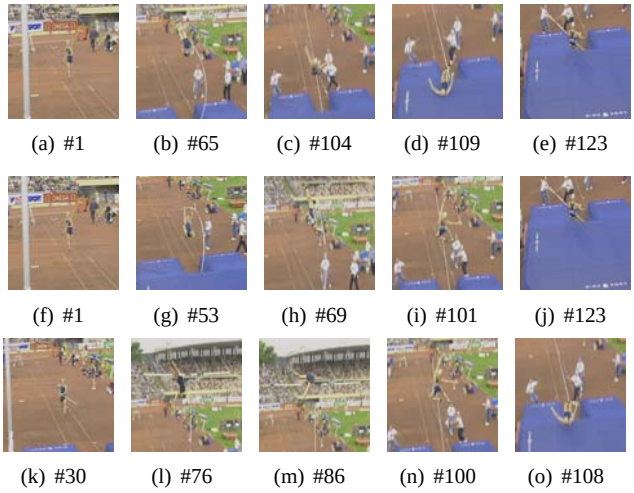


Fig. 2. The estimated solution (black circles) and the L_l curves are projected on $d(x, y)$ for pole vault (left) and showing (right) shots.

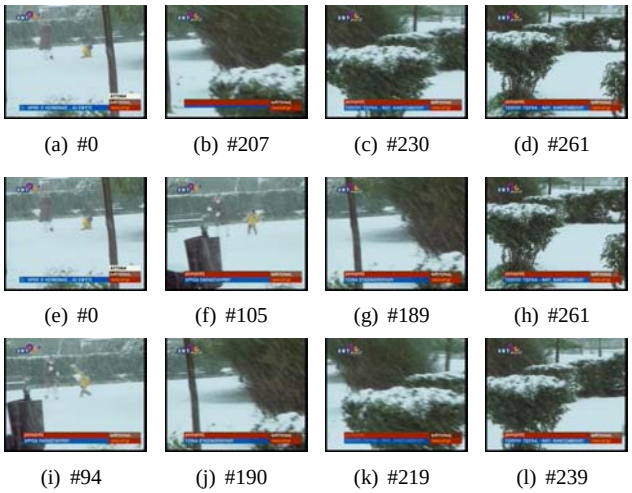


(a) #1 (b) #65 (c) #104 (d) #109 (e) #123

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

(k) #30 (l) #76 (m) #86 (n) #100 (o) #108

Fig. 3. Results of Iso-Content Distance, Iso-Content Distortion and Iso-Content MINMAX schemata in pole vault shot using five key frames. **(a),** $\cdots$**, (e)** The selected key frames under Iso-Content Distance principle. **(f),** $\cdots$**, (j)** The selected key frames under Iso-Content Distortion principle. **(k),** $\cdots$**, (o)** The selected key frames under Iso-Content MINMAX principle.

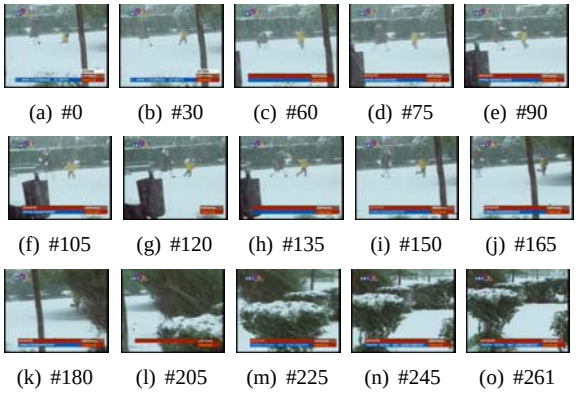


(a) #0 (b) #207 (c) #230 (d) #261

(e) #0 (f) #105 (g) #189 (h) #261

(i) #94 (j) #190 (k) #219 (l) #239

Fig. 4. Results of Iso-Content Distance, Iso-Content Distortion and Iso-MINMAX schemata in snowing shot using five key frames. **(a),** $\cdots$**, (d)** The selected key frames under Iso-Content Distance principle. **(e),** $\cdots$**, (h)** The selected key frames under Iso-Content Distortion principle. **(i),** $\cdots$**, (l)** The selected key frames under Iso-Content MINMAX principle.



(a) #0 (b) #30 (c) #60 (d) #75 (e) #90

(f) #105 (g) #120 (h) #135 (i) #150 (j) #165

(k) #180 (l) #205 (m) #225 (n) #245 (o) #261

Fig. 6. The snowing sequence which contains 262 frames.

[2] J. C.-M. Lee and D. M.-C. Ip, "A robust approach for camera break detection in color video sequence," in *IAPR Workshop on Machine Vision Application*, 1994.

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

[4] 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.

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

[6] Y. Avrithis, A. Doulamis, N. Doulamis, 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.

[7] Z. Li, G. Schuster, and A. Katsaggelos, "Minmax optimal video summarization," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 15, no. 10, pp. 1245 – 1256, 2005.

[8] Y.-F. Ma, X.-S. Hua, L. Lu, and H. Zhang, "A generic framework of user attention model and its application in video summarization." *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[9] C. Panagiotakis, G. Georgakopoulos, and G. Tziritas, "The curve equipartition problem," *submitted to Computer Aided Geometric Design*, 2007. [Online]. Available: http://www.csd.uoc.gr/˜cpanag/papers/EPcagd07.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] C. Panagiotakis, A. Doulamis, and G. Tziritas, "Equivalent key frames selection based on iso-content distance and iso-distortion principles," in *8th Inter. Workshop on Image An. for Mult. Inter. Serv. (accepted)*, 2007.

[13] 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.