21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Key Frame Selection Based on Jensen-Rényi Divergence

Qing Xu, Xiu Li, Zhen Yang, Jie Wang, Mateu Sbert*, Jianfu Li+
School of Computer Science and Technology, Tianjin University, 300072 Tianjin, China
**Graphics and Imaging Laboratory, Universitat de Girona, 17071 Girona, Spain*
+*College of Computer Science and Technology, CAUC, 300300 Tianjin, China*
Email:qingxu.itcn@gmail.com

Abstract

The key frame extraction is designed for obtaining a (very) compressed set of video frames that summarizes the essential content of a video sequence. In this paper, a well-known information theoretic measure, the Jensen-Rényi divergence (JRD), is studied to estimate the frame-by-frame distance between consecutive video images, for segmenting shots/subshots and for choosing key frames. Our new key frame extraction method, which is effective and computationally fast, contributes to a good and quick understanding of a large amount of video data.

1. Introduction

Digital videos have become more and more popular in our current society, thanks to the development of video capture devices. It is particularly required to have a good and quick understanding of general video sequences in a lot of real applications. But by and large, this task is not trivial, due to the fact that understanding videos means essentially dealing with a huge amount of data. Fortunately, video key frames, which are a set of (very) compact images representing the main content of the original video data, can be utilized for this purpose. In a word, key frame extraction is a quick and sound way for summarizing a video sequence.

In this paper, we propose a generic and effective method for video key frame selection. The novel technique we develop is based on a powerful information theoretic tool, namely the well-known *JRD* [17], which is applied as a difference measurement between two consecutive video frames for recognizing shot/subshot boundaries and for extracting key frames. The reason for making use of this measure from information theory is that, the theory itself is so general and powerful that it has been widely accepted in quite a lot of areas, such as computer vision [5]. As a matter of fact, quite many key frame selection algorithms have been invented in the literature [15][11]. Importantly enough, the classic information theory metrics, including *Shannon* entropy and *Mutual Information* (*MI*) [4], have been recently demonstrated their good capability of calculating similarity/difference between video frames for key frame selection in any kind of video sequence [10] [2]. A latest technique, based on using the *Jensen-Shannon divergence* (*JSD*), has been advanced for key frame extraction [16], behaving better than *Shannon* entropy and *MI* based approaches [10] [2]. Notably, our new method achieves improved performance when compared with that employing *JSD* [16].

The rest of this paper is structured as follows. Related work is briefly described in section 2. Our key frame selection technique is detailed in section 3. In section 4, experimental results are presented and discussed. Finally section 5 concludes the paper, and some future work is also mentioned.

2. Related Work

For the sake of this paper, we review the key frame selection methods that specifically use the information theoretic measures [10][2][6][12][16]. Mentzelopoulos and Psarrou [10] choose a key frame by counting a sufficient distance of Shannon entropy from the current one. Černeková et al. [2] apply *MI* to measure the similarity of the consecutive video frames for detecting shot boundaries and then for extracting key frames within shots. Janvier et al. [6] reduce the key frame selection to obtaining a solution to a cost function, established by using the *Jeffrey* divergence and a Minimum Message Length criterion in information theory. Omidyeganeh et al. [12] employ the frame-by-frame distance eval-

Thank NSFC (61179067, 61103005, 60879003), Spanish and Catalan Government (TIN2010-21089-C03-01, 2009-SGR-643 and 2010-CONE2-00053) for funding.

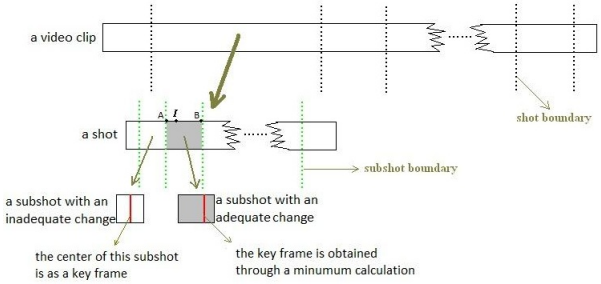

Figure 1. The computational mechanism for our key frame selection

uated from the *Kullback-Leibler* divergence on Generalized Gaussian Density parameters of wavelet coefficients of video images for separating shots/clusters and for obtaining the key frames. In our earlier work [16], *JSD* is utilized to divide a video into shots/subshots and then to choose key frames.

3. Key Frame Extraction Method

3.1 The Core Computational Mechanism

Usually, a general video sequence can be structured in a hierachical way; that is, a video (clip) can be divided into shots and then into video frames [7]. Accordingly the basic computational mechanism of our key frame selection method is to segment a video clip into different shots, then into subshots, and to choose key frames. The basic idea driving the segmentation of a video clip into shots/subshots is based on estimating the difference between consecutive video frames: the *JRD* is here proposed as the metric for this purpose.

According to [17], a *JRD* between video frames f_{i-1} and f_i can be obtained:

$$JRD(f_{i-1}, f_i) = R_q(\frac{p_{f_{i-1}} + p_{f_i}}{2}) - \frac{R_q(p_{f_{i-1}}) + R_q(p_{f_i})}{2}, \quad (1)$$

where $R_q(p) = \frac{1}{1-q} log \sum_{i=1}^{n} p_i^q (0 < q < 1)$ is a concave version of *Rényi* entropy [13], $p_{f_{i-1}}$ and p_{f_i} are the respective probability density functions of f_{i-1} and f_i, which are normalized from their intensity histogram distributions. Notice, the difference between two video images by *JRD* is actually a sum of the correspondences for the three RGB channels.

3.2 Locating Shots, Subshots and Key Frames

The entire computing procedure for our key frame selection is depicted in Figure 1.

For a video clip, all the *JRD* data are obtained by evaluating the *JRD* for each pair of two consecutive video frames. Because a shot boundary reveals an abrupt variation between two consecutive video frames, we locate the shot boundaries by detecting the spikes at the *JRD* data. If $\frac{JRD(f_{i-1}, f_i)}{JRD_w(f_{i-1}, f_i)} \geq \delta^*$ then a shot boundary is identified, where $JRD_w(f_{i-1}, f_i)$ is an average of the $JRD(f_{i-1}, f_i)$ neighbors on a temporal window with a size of w (in this paper, $w = 5$), and $\delta^* = 2.6$ is the threshold experimentally defined.

Within a shot, however, some adequate content variations could appear, for instance, possibly due to gradual scene transitions. Apparently, the video content variation can be characterized by the gradient of *JRD*, $\Delta(f_i) = JRD_w(f_i, f_{i+1}) - JRD_w(f_{i-1}, f_i)$, which is the rate at which a *JRD* value changes relative to change over time. In practice, a window-sized version of *JRD* gradient, $\Delta_w(f_i)$, is employed to filter out some possible minor perturbations of gradient data. If $|\Delta_w(f_I)| \geq \Delta_w^*$ (in this paper Δ_w^* is experimentally determined as 1.5×10^{-3}), then an adequate content change inside a shot is detected at the video frame f_I. Starting from the outlier f_I, its temporally closest left and right frames f_A and f_B within this shot are obtained satisfying

$$A = max\{i || \Delta_w(f_i)| \leq \nabla_w^* \wedge i < I\}, \quad (2)$$

$$B = min\{i || \Delta_w(f_i)| \leq \nabla_w^* \wedge i > I\}, \quad (3)$$

here ∇_w^* is a preestablished threshold, 5×10^{-6}. Now, the video section $[A, B]$ is thus a "complete" subshot with an adequate content variation. In this way each video section with an adequate content variation is identified as a subshot, and thus a shot is divided into consecutive subshots with the borders of all the subshots with adequate content variations. If two subshots both with adequate changes are temporally close enough (namely the difference between the right border of the left subshot and the left border of the right subshot is not larger than a predefined threshold, here it is taken as 5), then the two subshots and the middle section are combined into a single subshot with an adequate content change. Note that each subshot is either with an adequate content variation or with not.

A key frame is extracted, for a subshot, depending on the content variation. That is, the center frame of a subshot with an inadequate variation is deployed as a key frame. The key frame is obtained, for a subshot with an adequate variation, through minimizing the summed *JRD* values between it and all the others.

4. Experimental Results

In our experiment, we compare the proposed algorithm based on *JRD* with the *JSD* driven method [16]. All the parameters used for the *JSD* based technique are strictly according to the setting defined in [16]. Good entropic index for *JRD*, set by experimentation, uses values in $(0.2, 0.6)$, and is taken as 0.4 in this paper. *Uniform Sampling* (*UF*) [15], which is the

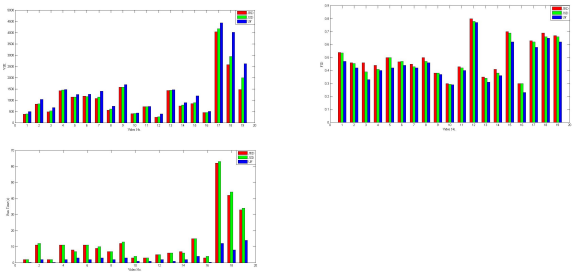

Figure 2. The *VSE*, *FID* and run time by different algorithms on the test videos

most straightforward and fastest key frame extraction method, is employed as a baseline, also for comparison.

We have done extensive tests for the three key frame selection methods on a large amount of video sequences. The test videos, including object motion, camera moving, panning, wipes, fade in/out, zoom in/out as well as some static scenes, are obtained from the web site "The Open Video Project" [1]. Table 1 shows the main information of each test video, such as the time duration (in seconds) and number of video frames. All the experiments are done based on a Windows PC with Intel Core i5 2.53 GHz CPU and 2GB RAM.

Table 1. Test videos

Video No.	Video Name	Length (s)	No. of Frames	Video No.	Video Name	Length (s)	No. of Frames
1	0037	27	830	11	Industry	36	1079
2	160	50	1512	12	NASAKSN-Shut	30	928
3	1234	28	854	13	senses100	58	1747
4	BOR04_002	77	2315	14	UGS06_006	40	1226
5	BOR06_002	65	1979	15	UGS08_016	87	2618
6	BOR06_004	62	1886	16	UGS13_005	25	776
7	BOR08_007	58	1759	17	cscw00_02_m4	399	8377
8	BOR14_001	36	1083	18	cscw92_05_m4	239	7159
9	BOR19_007	74	2219	19	NASAWF-GIFTS	185	5571
10	hcil2004_01_m1	36	921				

Two widely used quantitative measures, Video Sampling Error (*VSE*) [8] and Fidelity (*FID*) [3], are deployed for performance evaluation on different key frame extraction methods. The similarity of two images, which is required for the computation of *VSE* and *FID*, is obtained by the second scheme in a well-accepted literature [14]. A low *VSE* and/or a high *FID* is an indicator for a good performance on key frame selection, and *vice versa*. In Figure 2, the *VSE*, *FID* and run time values, which are resulted from the different algorithms, are displayed. It is clear that, all in all, *JRD* performs better than *JSD*. We believe that this is due to the better entropic ability of *JRD* to assess the difference among the probability distributions.

Figures 3 - 6 provide several examples to show the better performance by *JRD*. The *JRD* and *JSD* plots for the two test videos "UGS06_006" and "BOR14_001" are given in Figures 3 and 5: we use a black dash-dotted line, a green dashed line and a red point to respectively represent a shot boundary, a subshot separation and a key frame. Correspondingly, the key frames extracted are shown in Figures 4 and 6.

Figures 3 and 4 exhibit the behaviors of different key

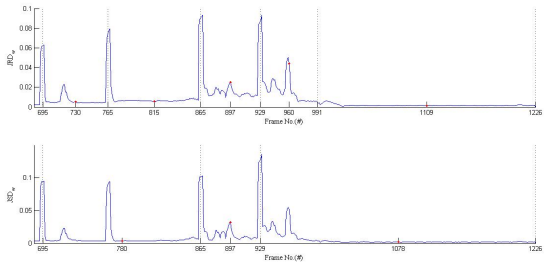

Figure 3. The *JRD* and *JSD* plots for the test video "UGS06_006"

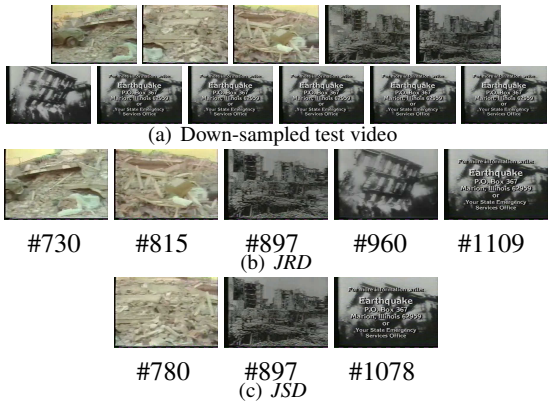

Figure 4. Comparison of different methods on the test video "UGS06_006"

frame selection methods, for a clip of "UGS06_006". This video clip is with big content changes, including a sufficient camera motion (#695 - #865) and panning (#866 - #929), a large object moving (#930 - #991) and a static scene with a caption that is being generated (#992 - #1226). Apparently, *JRD* indicates a shot boundary at #765, thus two key frames #730 and #815 are extracted and they can satisfactorily point out that the camera is moving. In addition, the key frame #960, resulted from the identification of a shot cut at #991 by *JRD*, clearly renders an object motion. Unfortunately, the key frames selected by *JSD* cannot well depict the camera and object motions. In sum, our *JRD* based technique does very well for identifying hard cuts and for locating key frames, and can better represent the video contents compared with those by *JSD*.

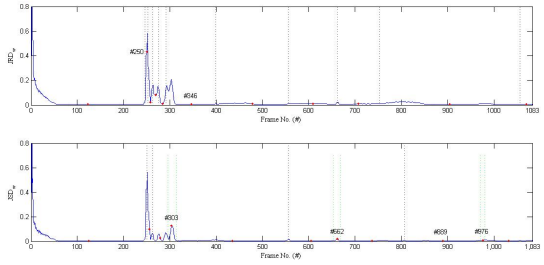

Figure 5. The *JRD* and *JSD* plots for the test video "BOR14_001"

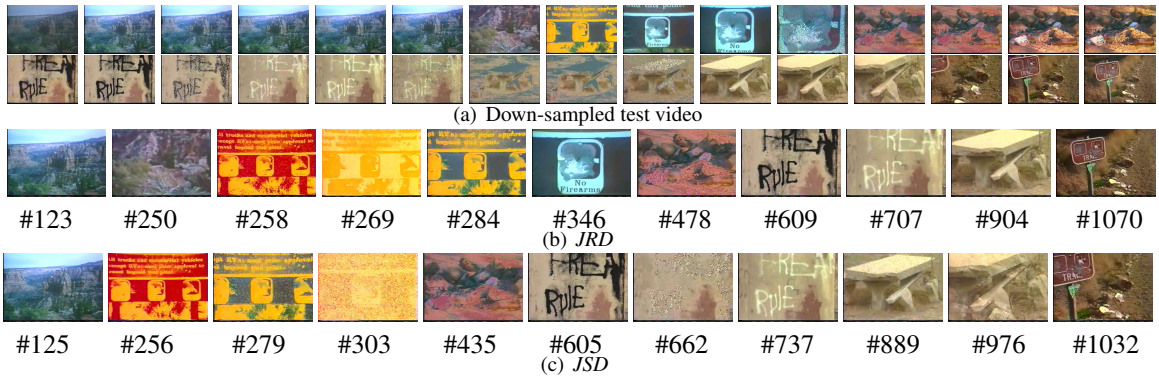

Figure 6. Comparison of different methods on the test video "BOR14_001"

Figures 5 and 6 present that *JRD* achieves superior results than *JSD* on "BOR14_001", which is with abundant scene switches and is a typical general video sequence in many real applications. *JRD* can extract the key frames indicating scene switches (#250 and #346), and this obviously benefits the understanding of the original video. The key frames by *JRD* have less redundancy, however *JSD* selects the duplicated #889 and #976 as the key frames. As for the image quality of the extracted key frames themselves, while the outputs by *JSD* can be unsatisfying (#303 and #662), the products by *JRD* are usually acceptable, and apparently this is an advantage for a key frame selection technique.

5. Conclusion and Future Work

A novel and effective key frame selection method has been presented. *JRD* has been demonstrated to perform better than *JSD*, by our shot-based computational mechanism. For the future work, we are going to utilize the wavelet representation of video frames for the key frame extraction, because the wavelet transform of an image profits the human perception on its content [9]. Also we will consider the motion within the video to select key frames.

References

[1] http://www.open-video.org/index.php.

[2] Z. Cernekova, I. Pitas, and C. Nikou. Information theory-based shot cut/fade detection and video summarization. *IEEE Trans Circuits Syst video technol*, 16(1):82–91, January 2006.

[3] H. S. Chang, S. Sull, and S. U. Lee. Efficient video indexing scheme for content-based retrieval. *IEEE Trans Circ Syst Video Technol*, 9(8):1269–1279, December 1999.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd Ed.* San Francisco: Wiley-Interscience, 2006.

[5] F. Escolano, P. Suau, and B. Bonev. *Information Theory in Computer Vision and Pattern Recognition*. Springer London, 2009.

[6] B. Janvier, E. Bruno, T. Pun, and S. Marchand-Maillet. Information-theoretic temporal segmentation of video and applications: multiscale keyframes selection and shot boundaries detection. *Multimed Tools Appl*, 30(3):273–288, 2006.

[7] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Video abstracting. *Commun ACM*, 40(12):54–62, December 1997.

[8] T.-C. Liu and J. R. Kender. Computational approaches to temporal sampling of video sequences. *ACM T. Multim. Comput.*, 3(2):217–218, 2007.

[9] S. Mallat. *A wavelet tour of signal processing, 2nd Ed.* San Diego: Academic Press, 1999.

[10] M. Mentzelopoulos and A. Psarrou. Key-frame extraction algorithm using entropy difference. In *Proc. ACM SIGMM Int. Conf. Workshop Multimedia Information Retrieval*, pages 39–45, 2004.

[11] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *J Vis Commun Image Represent*, 19(2):121–143, February 2008.

[12] M. Omidyeganeh, S. Ghaemmaghami, and S. Shirmohammadi. Video keyframe analysis using a segment-based statistical metric in a visually sensitive parametric space. *IEEE Trans. Iamge Process.*, 20(10):2730–2737, October 2011.

[13] A. Rényi. On measures of entropy and information. In *Selected papers of Alfréd Rényi*, pages 525–580, 1976.

[14] M. A. Stricker and M. Orengo. Similarity of color images. *Proc SPIE*, 2420(2):381–392, May 1995.

[15] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM T. Multim. Comput.*, 3(1):1–37, 2007.

[16] Q. Xu, P.-Ch.Wang, B. Long, M. Sbert, M. Feixas, and R. Scopigno. Selection and 3d visualization of video key frames. In *Proceedings of IEEE International Conference on Systems Man and Cybernetics (SMC)*, pages 52–59, 2010.

[17] H. Yun, A. B. Hamza, and H. Krim. A generalized divergence measure for robust image registration. *IEEE Trans Signal Process*, 51(5):1211–1220, May 2003.