F-DIVERGENCES DRIVEN VIDEO KEY FRAME EXTRACTION

Xiaoxiao Luo[a], *Qing Xu*[a*], *Mateu Sbert*[b], *Klaus Schoeffmann*[c]

[a]School of Computer Science and Technology, Tianjin University, 300072 Tianjin, China
[b]Graphics and Imaging Laboratory, Universitat de Girona, 17071 Girona, Spain
[c]Alpen-Adria-Universität Klagenfurt, Universitätsstr. 65-67, 9020 Klagenfurt, Austria
Email:qingxu@tju.edu.cn

ABSTRACT

This paper proposes a shot-based key frame selection technique, aiming at generating a condensed set of frames representing the essential content of a video sequence. Inspired by the successful utilization of *Jensen-Shannon Divergence* (*JSD*) and *Jensen-Rényi Divergence* (*JRD*) for key frame selection [1] [2], we investigate several popularly accepted f-divergences for calculating the frame-by-frame distance to segment the video clip and then to obtain the key frames. Based on simulation and real test videos, the performances of the key frame selection method by using different versions of f-divergences are systematically analyzed. Extensive experimentation shows that, compared with the methods using *JSD* and *JRD*, the new technique is slightly better and computationally faster.

Index Terms— key frame selection, f-divergences, *Jensen-Shannon Divergence* (*JSD*), *Jensen-Rényi Divergence* (*JRD*)

1. INTRODUCTION

Digital videos are becoming more and more critical in many application fields, such as content-based video analysis, video skimming and retrieval, due to the popularity of video capture devices. Especially, with the development of interactive digital TV and with the explosion of Internet videos, a fast and good understanding of video clips of any kind is highly demanded. Key frames of a video, which are the reduced set of still images depicting the visual content of the original video data, keep us from the heavy load of the large video data and meanwhile, provide us with a quick grasp of the original video contents.

For the sake of dealing with video clips of any kind, the shot-based key frame selection approaches through using the general information theoretic measures, such as *Jensen-Shannon Divergence* (*JSD*) and *Jensen-Rényi Divergence* (*JRD*), have been presented with very good results [1] [2]. However, the weighted average of probability distributions is involved in the calculation of *JSD* and *JRD*, which requires high computational costs. Apparently this is not so good for the fast video key frame selection, which is critically required in many current applications such as the browsing of mobile videos [3] [4]. Motivated by employing the divergences that are more efficiently computed, we propose to exploit several widely used f-divergences (*FD*s) for shot-based key frame selection. In total, four versions of *FD*s, exactly speaking, Chi-square distance (χ^2) and square root of Chi-square distance (χ), Hellinger distance (h^2) and square root of Hellinger distance (h), are studied in depth in this paper.

It is undoubtedly meaningful for a user to choose which kind of *FD*s to be applied in the key frame selection for his specific purpose in real practice. Generally real videos include both abrupt and gradual scene changes. In fact, the videos just with abrupt scene changes are relatively easy for key frame selection. However, the handling of video data with gradual scene changes, usually including several possible cases such as dissolve, fade in/out, zoom in/out, object and camera motions, is more difficult [5] [6]. So we specially devise a set of simulation test videos with gradual transitions and, each simulation video contains only one case mentioned above, in order to clearly evaluate the detailed performances by different versions of *FD*s used for key frame selection. Extensive experimentation additionally indicates our proposed shot-based computational mechanism using *FD* is efficient and also effective.

The remainder of this paper is organized as follows. In the next section, related work for key frame extraction algorithms is reviewed. Our key frame selection technique is detailed in section 3. Section 4 describes the simulation videos we designed and experimental results. The final section 5 concludes the paper.

2. RELATED WORK

For the sake of this paper, the key frame selection techniques using information theoretic measures are reviewed. Černeková et al. [6] propose a method based on mutual information (MI) and joint entropy (JE) to evaluate the simi-

Qing Xu is the corresponding author.

To appear in Proceedings of ICME 2014, July 14-18, 2014, Chengdu, China, (c) IEEE 2014

larity between contiguous video frames. Abrupt cut and fade boundaries are respectively detected where the MI and JE are abruptly decreased, and then key frames are selected from each shot. Mentzelopoulos and Psarrou [7] employ the Shannon entropy of the image probability distribution. The distance of Shannon entropy is computed between the last defined key frame and the current frame, which will be taken as a new key frame if this distance is sufficient enough. Omidyeganeh et al. [8] use the coefficients of the wavelet transform subbands to obtain the features of video frames. Based on these features, the *Kullback-Leibler* distance is employed as the similarity measure to segment a video into shots and clusters. For each cluster, the video frame, which is closest to all its neighboring frames in this cluster and differs from the frames outside this cluster, is selected as the key frame. In [1] and [2], *JSD* and *JRD* are respectively utilized as the metrics to calculate the difference of video frames to identify the shots and subshots, and then to obtain the key frames.

3. *F*-DIVERGENCES KEY FRAME EXTRACTION METHOD

A video can be considered as a succession of shots, each representing an event or continuous success of actions [9]. In our approach, the video stream is divided into non-overlapping shots and then into subshots, and one frame is chosen for each subshot. The f-divergence is here provide us the metric for the distance of consecutive frames.

3.1. *F*-divergences distance measure

Given a convex function f: $[0,\infty) \rightarrow \mathbb{R}$, the f-divergence of the probability distributions $p = \{p_1, p_2 ... p_n\}$ and $q = \{q_1, q_2 ... q_n\}$ is given by

$$FD(p||q) \equiv \sum\nolimits_{i=0}^{n} q_i f(\frac{p_i}{q_i}). \quad (1)$$

The f-divergence is introduced by Csiszár [10] and Ali & Silvey [11] as a measure of discrimination between two probability distributions.

Coinciding with different convex functions f, some particular cases of f-divergences [12] between the ith frame and (i+1)th frame in a video are defined. In the following, we take $x > 0$.

- Chi-square distance [13]
  If $f(x) = (x-1)^2$, the Chi-square distance is given by

$$\chi^2(f_i, f_{i+1}) = \sum\nolimits_{j=0}^{n} \frac{(p_{i,j} - p_{i+1,j})^2}{p_{i+1,j}}; \quad (2)$$

- Square root of Chi-square distance [14]

$$\chi(f_i, f_{i+1}) = \sqrt{\sum\nolimits_{j=0}^{n} \frac{(p_{i,j} - p_{i+1,j})^2}{p_{i+1,j}}}; \quad (3)$$

- Hellinger distance [15]
  If $f(x) = \frac{1}{2}(1-\sqrt{x})^2$, the Hellinger distance is given by

$$h^2(f_i, f_{i+1}) = \frac{\sum_{j=0}^{n} (\sqrt{p_{i,j}} - \sqrt{p_{i+1,j}})^2}{2}; \quad (4)$$

- Square root of Hellinger distance [16]

$$h(f_i, f_{i+1}) = \sqrt{\frac{\sum_{j=0}^{n} (\sqrt{p_{i,j}} - \sqrt{p_{i+1,j}})^2}{2}}, \quad (5)$$

where $\{p_{i,0}, p_{i,1}, \cdots, p_{i,n}\}$ is the probability distribution of gray level, corresponding to the normalized intensity histogram distribution of the ith frame, assuming that the video gray levels vary form 0 to n. Notice, the distance between frames is the sum of correspondences for the three RGB channels. $FD(f_i||f_{i+1})$ represents $\chi^2(f_i, f_{i+1})$, $\chi(f_i, f_{i+1})$, $h^2(f_i, f_{i+1})$ or $h(f_i, f_{i+1})$.

3.2. Key frame extraction method

In our approach, the *FD*s between each pair of two consecutive frames are calculated by using (2)-(5). A spike of *FD* indicates the existence of a shot boundary. In order to locate spikes, a ratio is defined as $\delta = \frac{FD}{FD_w}$, where FD_w is the local mean of *FD* on a w size window ($n_w = 5$ in this paper). A shot boundary is identified where δ is greater than a predetermined threshold δ^*.

Next, each shot is divided into several subshots (note that if the content change of a video shot is small then this shot does not need to be divided into subshots). Within a shot, a significant content change, which is probably caused by a notable object/camera motion, could occur. The gradient of FD_w, calculated as $\Delta(i) = FD_w(f_i||f_{i+1}) - FD_w(f_{i-1}||f_i)$, is employed to detect the significant content change. Considering that some minor perturbations of Δ could exist, the local mean of Δ on a window, noted as Δ_w, is actually used for this sake. If $|\Delta_w(i)|$ is greater than a pre-defined threshold Δ_w^*, then an obvious content change inside the shot appears around the frame f_i. The left and right closest frames to f_i, f_a and f_b, which satisfy $\Delta_w(f_a) \simeq 0$ and $\Delta_w(f_b) \simeq 0$, are respectively identified as the beginning and end of the significant content change. In fact, f_a and f_b are respectively the left and right boundaries of a subshot $[f_a, f_b]$ with the significant content change. For a subshot with the significant content change, the frame being most similar to all the others in this subshot is selected as the key frame of the subshot. For a subshot without the significant content change, the center frame of it is extracted as a key frame. The procedure is detailed in pseudo code (see details at algorithm 1). In this paper, $\delta^* = 2.6$, $\Delta_w^* = 1.5 \times 10^{-3}$, $\nabla_w^* = 5 \times 10^{-6}$ are used for the pre-established thresholds.

Algorithm 1 shot based key frame selection

Input: A video of N frames

Output: Key frames

for each frame f_i $(1 < i \leq N)$ **do**
  Compute $FD(f_{i-1}||f_i)$;
  Compute $FD_w(f_{i-1}||f_i)$
  $FD_w(f_{i-1}||f_i) = \frac{1}{n_w} \sum_{j=i-\lfloor \frac{n_w}{2} \rfloor}^{i+\lfloor \frac{n_w}{2} \rfloor} FD(f_{j-1}||f_j)$;
  Compute $\delta(i-1,i) = \frac{FD(f_{i-1}||f_i)}{FD_w(f_{i-1}||f_i)}$;
  if $\delta(i-1,i) > \delta^*$ **then**
    A shot boundary between f_{i-1}, f_i is identified;
  end if
end for
for each shot **do**
  for each frame f_i in the shot **do**
    Compute $\Delta(i) = FD_w(f_i||f_{i+1}) - FD_w(f_{i-1}||f_i)$;
    Compute $\Delta_w(i) = \frac{1}{n_w} \sum_{j=i-\lfloor \frac{n_w}{2} \rfloor}^{i+\lfloor \frac{n_w}{2} \rfloor} \Delta(j)$;
    if $|\Delta_w(i)| > \Delta_w^*$ **then**
      $a = max\{j||\Delta_w(f_j)| \leq \nabla_w^* \wedge i < j\}$;
      $b = min\{j||\Delta_w(f_j)| \leq \nabla_w^* \wedge i > j\}$;
    end if
  end for
end for
for each subshot with significant content change **do**
  Locate the key frame f_k of the subshot $[f_a, f_b]$
  $k = argmin\{a < k < b| \sum_{j=a}^{b} FD(f_k||f_j)\}$;
end for
for each subshot without significant content change **do**
  Select the center frame as key frame;
end for

4. EXPERIMENTS ON SIMULATION AND REAL TEST VIDEOS

The proposed key frame selection algorithm by *FD*s is compared with those by *JSD* and *JRD*. Extensive tests have been performed based on simulation and real test videos. Simulation videos are specially designed for gradual transitions. And the real test videos, including films, sports, advertisements, news and so on, are obtained from the web site "The Open Video Project" [17]. Generally a good entropic index used for *JRD* takes a value in [0.4, 0.5): here 0.45 is used. Two widely used objective quantitative measures, Video Sampling Error (*VSE*) [18] and Fidelity (*FID*) [19], are used to evaluate the performances of key frame extraction methods. The lower the *VSE*, the better the evaluated method is, and vice versa. The higher the *FID*, the better the evaluated method is, and vice versa. All the experimental results are obtained based on a Windows PC with Intel Core i7 2.0 GHz CPU and 16GB RAM.

Table 1 lists the average runtime comparisons. The best values are on bold. It is obvious that the methods by *FD*s get higher efficiency, and this is due to the fact that *FD*s can be computed faster than *JSD* and *JRD*.

Table 1. Average runtime by each method (seconds)

Duration of Test Videos \ Methods	χ^2	χ	h^2	h	*JRD*	*JSD*
less than 30 seconds	1	1	**0.7**	1.5	1.7	2
30 seconds to 1 minutes	**2.5**	3	**2.5**	3	4.6	5
3 to 4 minutes	39	37.5	**36.6**	38.8	41	48
6 to 7 minutes	**46**	50	53	56	71	65
14 to 15 minutes	168	172	**162**	174	218	212

4.1. Experimental results on simulation videos

For testing and analyzing the key frame selection algorithms on videos with gradual transitions, which are more difficult to be handled than those only with hard cut, we devise six types of simulation videos listed in Table 2. These simulation videos are representative of the gradual transitions, including "object motion", "camera motion", "object and camera motion", zoom, fade, and dissolve. It is well known that gradual transitions take place across a number of consecutive frames and, the speed of different transitions in videos varies a lot. So the speed of content change is an important feature of gradual transitions. In order to systematically analyze the proposed key frame selection technique based on four versions of *FD*s, each type of simulation video is designed as a number of concrete implementations with different speeds of content change. The design method is that, with the given number of key frames, a certain number of interpolation video frames between the each two given key frames are obtained. A large number of the interpolation video frames result in an implementation of simulation video with a slow speed of content change. On the contrary, a small number of the interpolation video frames lead to an implementation of simulation video with a fast speed of content change. For the sake of this paper, we define a metric "transition-rate" to represent the speed of content change for the simulation video, as follows:

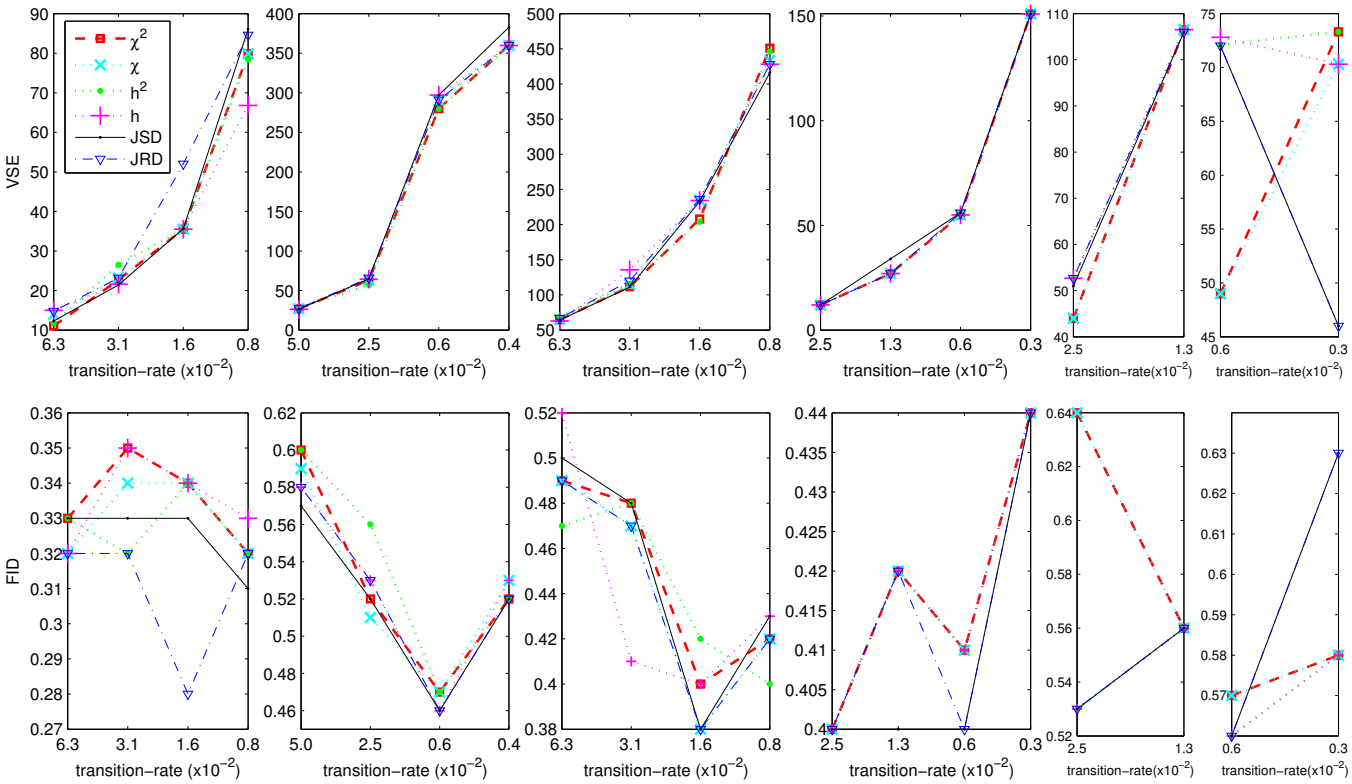
$$\text{transition-rate} = \frac{\text{number of the given key frames}}{\text{number of frames}}. \quad (6)$$

An example of a fade type simulation video is shown in Figure 3. Here the fade transition starts from a normal illuminated frame to an overexposed image, and continues to another frame with normal illumination. The two given key frames are the first and last frames of this fade.

Table 2. Type of simulation videos

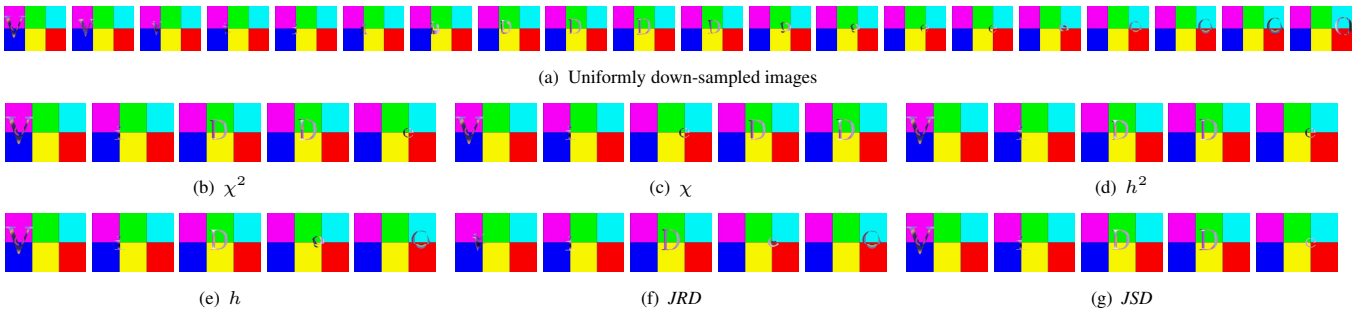
object motion	zoom
camera motion	fade
object and camera motion	dissolve

Due to space limitations, only some of the *VSE* and *FID* values by different methods on the simulation videos are exemplified in Figure 1. On the abscissa are the transition-rates



(a) Object motion (b) Camera motion (c) Object and camera motion (d) Zoom (e) Fade (f) Dissolve

Fig. 1. The *FID* and *VSE* by different algorithms on the simulate videos



(a) Uniformly down-sampled images

(b) χ^2 (c) χ (d) h^2

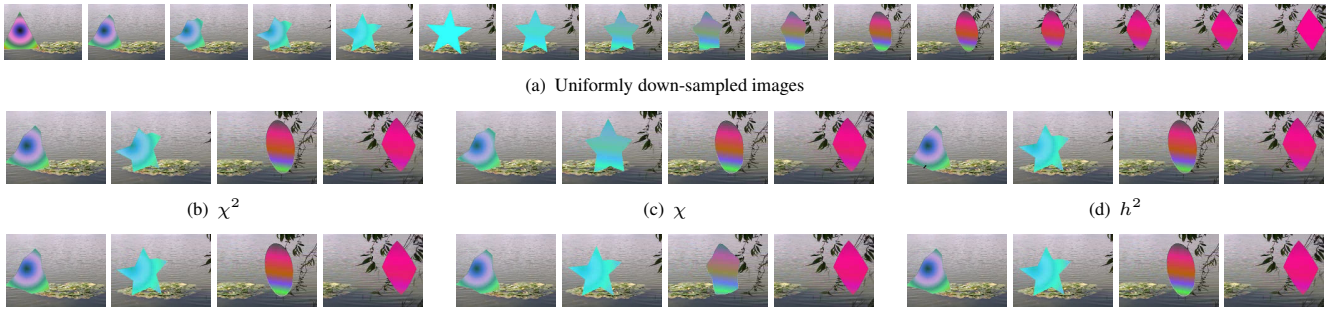
(e) h (f) *JRD* (g) *JSD*

Fig. 2. Comparison of different methods on "ViDeO4"

Fig. 3. An example of a fade type simulation video

of each video. Figures 1(a), 1(b), 1(d), and 1(e) point out that *FD*s achieve superior results than *JSD* and *JRD* for videos in types of "object motion", "camera motion", zoom and fade. In the case of "object and camera motion", the *VSE* and *FID* by *FD*s are comparable with those by *JSD* and *JRD*. The performances by *FD*s for videos with dissolves are inferior to those by *JSD* and *JRD*. As a result, we can speak that, in general, the performances of *FD*s based key frame selection methods are better than those of *JSD* and *JRD* based ones, for videos with gradual transitions. As for the differences between the schemes based on χ, χ^2, h and h^2, χ^2 and h^2 behave superior to χ and h, for videos with large content variations (in the high transition-rate). Whereas χ and h outperform χ^2 and h^2, for videos with the low transition-rate.

Figure 2 exhibits the key frames selected by different methods, for a "object motion" video clip "ViDeO4". The transition-rate of "ViDeO4" is 0.781×10^{-2}, which means that the object motion is relatively slow. In this video, five letters, V, i, D, e and O, come out in sequence with gradual appearance transformations. The given key frames are the five



(a) Uniformly down-sampled images

(b) χ^2 (c) χ (d) h^2

(e) h (f) *JRD* (g) *JSD*

Fig. 4. Comparison of different methods on "mCO1"

video images in which the five letters are exactly the V, i, D, e and O. It can be noticed that the measure h achieves superior results than the others, in terms of the complete video summarization with no duplication. The outputs by h correctly present these five letters. Note also that the results by *JRD* are acceptable, but the key frame for the letter "V" is not so good.

Figure 4 shows the key frames extracted from a "object and camera motion" video "mCO1". The transition-rate of "mCO1" is 5×10^{-2}, which means that the motion is fast. In the video, a 2D object gradually transforms in the shapes of triangle, star, oval and diamond, and in the meantime the camera pans. The outputs by *FD*s can represent the video content very well, and have no redundancy at all. However the third key frame by *JRD* cannot present a complete shape.

4.2. Experimental results on real test videos

Table 3 shows the *VSE* and *FID* results from the six methods. Clearly, *FD*s based methods outperform *JSD* and *JRD* with real test videos.

A few of key frames selected from a clip of a movie with large video content variations are shown in Figure 5. The key frame selection algorithms based on four versions of *FD*s, *JSD* and *JRD* obtain the same output, which can represent the video very well and have no redundancy at all. The methods based on *FD*s run on average 22% faster than those based on *JSD* and *JRD*.

(a) Uniformly down-sampled images

(b) *FD*s, *JSD*, and *JRD*

Fig. 5. The key frames selected from a clip of "Life of Pi"

Table 3. Measures by Different Methods

Video Name (No. of Frames)	Method	*VSE*	*FID*
Lemon (739)	χ^2	**64**	**0.617**
	χ	66	0.614
	h^2	**64**	0.616
	h	66	0.613
	JSD	79	0.613
	JRD	75	**0.617**
Football (752)	χ^2	469	0.417
	χ	478	0.405
	h^2	477	0.402
	h	**467**	**0.433**
	JSD	476	0.390
	JRD	473	0.396
Metal (2882)	χ^2	**1153**	0.678
	χ	1164	**0.696**
	h^2	1180	0.617
	h	1233	0.604
	JSD	1220	0.522
	JRD	1218	0.524
uist02 (5689)	χ^2	2961	0.555
	χ	2981	0.548
	h^2	**2902**	**0.556**
	h	2980	0.547
	JSD	2950	0.545
	JRD	2937	0.551
UGS03_0 (5992)	χ^2	3651	0.428
	χ	3581	**0.440**
	h^2	3679	0.420
	h	**3482**	0.438
	JSD	3607	0.429
	JRD	3603	0.430

5. CONCLUSION

A novel video key frame selection approach driven by f-divergences is proposed. The approach is applied to extensive experiments on simulation and real videos. In addition, a

thorough analysis of the performances on *FD*s for key frame selection is provided. Experimental results demonstrate that the key frame selection by *FD*s is slightly better than that by *JSD* and *JRD*. We also verify that the key frame selection by *FD*s can run faster than that by *JSD* and *JRD*, due to the faster computation of *FD*s for video frames and also to the shot-based computational mechanism presented in this paper.

6. ACKNOWLEDGEMENT

This work has been funded by Natural Science Foundation of China (61179067, 60879003, 61103005), and Spanish (TIN2010-21089-C03-01) and Catalan Governments (2009-SGR-643 and 2010-CONE2-00053).

7. REFERENCES

[1] 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)*, 2010, pp. 52–59.

[2] Qing Xu, Xiu Li, Zhen Yang, Jie Wang, Mateu Sbert, and Jianfu Li, "Key frame selection based on jensen-rényi divergence," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 1892–1895.

[3] Marco A Hudelist, Klaus Schoeffmann, and Laszlo Boeszoermenyi, "Mobile video browsing with a 3d filmstrip," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 2013, pp. 299–300.

[4] Marco A Hudelist, Klaus Schoeffmann, and Laszlo Boeszoermenyi, "Mobile video browsing with the thumbbrowser," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 405–406.

[5] Jian Zhou and Xiao-Ping Zhang, "Video shot boundary detection using independent component analysis," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2005, pp. 541–544.

[6] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans Circuits Syst video technol*, vol. 16, no. 1, pp. 82–91, January 2006.

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

[8] 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.*, vol. 20, no. 10, pp. 2730–2737, October 2011.

[9] Boon-Lock Yeo and Bede Liu, "Rapid scene analysis on compressed video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 5, no. 6, pp. 533–544, 1995.

[10] Imre Csiszár., "Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten," *Magyar. Tud. Akad. Mat. Kutato Int. Kozl*, vol. 8, pp. 85–108, 1963.

[11] SM Ali and Samuel D Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 131–142, 1966.

[12] Sever S Dragomir, "Some inequalities for the csiszár f-divergence," *Inequalities for Csiszr f-Divergence in Information Theory*, 2000.

[13] Karl Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[14] Dominik Maria Endres and Johannes E Schindelin, "A new metric for probability distributions," *Information Theory, IEEE Transactions on*, vol. 49, no. 7, pp. 1858–1860, 2003.

[15] I Csisz et al., "Neue begründung der theorie quadratischen formen von unendlichen vielen veränderlichen derlichen," *Journal fr Reine und Angewandte Mathematik.*, vol. (136), pp. 210–271, 1909.

[16] Flemming Topsoe, "Some inequalities for information divergence and related measures of discrimination," *Information Theory, IEEE Transactions on*, vol. 46, no. 4, pp. 1602–1609, 2000.

[17] "http://www.open-video.org/index.php," .

[18] T.-Ch. Liu and J. R. Kender, "Computational approaches to temporal sampling of video sequences," *ACM T. Multim. Comput.*, vol. 3, no. 2, pp. 217–218, 2007.

[19] H. S. Chang, S. Sull, and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," *IEEE Trans Circ Syst Video Technol*, vol. 9, no. 8, pp. 1269–1279, December 1999.