Performance Evaluation of Video Summaries Using Efficient Image Euclidean Distance

Sivapriyaa Kannappan(✉), Yonghuai Liu, and Bernard Paul Tiddeman

Department of Computer Science, Aberystwyth University,
Aberystwyth SY23 3DB, UK
{sik2,yyl,bpt}@aber.ac.uk

Abstract. Video summarization aims to manage video data by providing succinct representation of videos, however its evaluation is somewhat challenging. IMage Euclidean Distance (IMED) has been proposed for the measurement of the similarity of two images. Though it is effective and can tolerate the distortion and/or small movement of the objects, its computational complexity is high in the order of $O(n^2)$. This paper proposes an efficient method for evaluating the video summaries. It retrieves a set of matched frames between automatic summary and the ground truth summary through two way search, in which the similarity between two frames are measured using the Efficient IMED (EIMED), which considers neighboring pixels, rather than all the pixels in the frames. Experimental results based on a publicly accessible dataset has shown that the proposed method is effective in finding precise matches and usually discards the false ones, leading to a more objective measurement of the performance for various techniques.

1 Introduction

A video summary is defined as a sequence of still or moving pictures which provides a concise representation of the video content, while the essential message of the original video is preserved [1]. There are two basic types of video summaries [2]: *static video summary* and *dynamic video skimming*. The former consists of a set of key frames, whereas the latter consists of a set of shots extracted from the original video [3]. The key benefit of video skimming is that the content includes both audio and motion elements, which enhance both the emotions and the amount of information conveyed by the summary. On the other hand, as key frames are not restricted to timing and synchronization issues, it is more versatile compared to consecutive display of video skims [3]. Hence we focus on static video summaries.

Many video summarization techniques have been proposed in the past few years [3–6]. Nevertheless the evaluation of those video summaries are quite challenging due to the lack of an efficient evaluation method and the judgement of interestingness or importance of the contents is usually subjective and application dependent.

G. Bebis et al. (Eds.): ISVC 2016, Part II, LNCS 10073, pp. 33–42, 2016.
DOI: 10.1007/978-3-319-50832-0_4

According to Troung and Venkatesh [2], the current evaluation methods in video summarization can be classified into three distinct groups such as (i) Result description, (ii) Objective metrics and (iii) User studies. Meanwhile De Avila *et al.* [3] proposed a novel evaluation method called Comparison of User Summaries (CUS) where the video summary is built by a number of users from the sampled frames. Those user summaries act as a ground truth, which are compared with the automatic summaries obtained by various methods. However, evaluation of those video summaries are tricky and usually subjective in nature.

Video summary evaluation by De Avila *et al.* [3] and Mei *et al.* [7] used only color features based on Manhattan distance to measure the similarity between automatic summary (AT) and ground truth summary (GT), alternatively the evaluation by Mahmoud [8] and Mahmoud *et al.* [4] incorporates both color and texture features based on the Bhattacharya distance. The downside of using color feature is that two different images may have the same color histogram. If so, false frame matches will be established. The texture feature may help to overcome this shortcoming. Though color and texture features give more perceptual assessment of the quality of video summaries, it is computationally expensive and challenging in terms of how both the features can be combined. Thus existing techniques may detect similar frames incorrectly between AT and GT for performance measurement, which are crucial for the development of more precise and robust methods.

As a result, we propose a simple and efficient approach for video summary evaluation. This method retrieves a set of potential matches between AT and GT using a two-way search from AT to GT and then to AT again. Wang *et al.* proposed IMage Euclidean Distance (IMED) [9] which considers the spatial relationship between all the pixels. This is computationally inefficient and somewhat unnecessary, considering especially the case, that the movements of the objects in the neighboring frames are relatively small. Thus, we propose to improve the IMED through considering only the neighboring pixels, just like a kernel with a size, let's say 3×3, for example, leading to an Efficient IMED (EIMED). The EIMED is used to measure the similarity between two frames for our method.

The proposed technique is validated using a publicly accessible dataset. The experimental results show that neighboring pixels are usually sufficient for the measurement of the similarity of different frames and some state-of-the-art techniques do not perform as well as described in the literature. Such findings will be helpful for other researchers to gain more insights into the performance of the state-of-the-art and help them to develop more advanced techniques.

The main contributions of this paper are:

1. We propose a simple and efficient two-way evaluation method using EIMED which considers the spatial relationship between the neighboring pixels alone
2. A comparative study between different summarization techniques shows their true relative performance, which will be vital for other researchers to further investigate the techniques

The rest of this paper is organised as follows: the proposed evaluation method is detailed in Sect. 2; the experimental results are presented in Sect. 3; and finally conclusions are drawn in Sect. 4.

2 Proposed Evaluation Method

Though different video summarization techniques have been proposed in the literature, performance evaluation of those techniques is still challenging. In this paper, we propose an efficient two-way evaluation method based on EIMED which is explained in the following sections. The main idea of the two-way evaluation method were detailed in [10] in which the similarity between the frames are measured using EIMED. The major advantage of our two way evaluation method is that it does not need to set up any threshold for retrieving the number of matched frames and thus has an advantage of easy implementation.

The key terms used in this paper: Automatic Summary (AT) denotes extracted key frames from various summarization techniques, Ground Truth User Summary (GT) denotes different user summaries obtained from [3].

2.1 IMage Euclidean Distance (IMED)

An image with a size of $M \times N$ pixels can be written as a vector $x = \{x^1, x^2, ...x^{MN}\}$ according to the gray level of each pixel. The conventional Euclidean distance $d_E^2(x_1, x_2)$ between vectorized images x_1 and x_2 is defined as [9,11]:

$$d_E^2(x_1, x_2) = \sum_{k=1}^{MN} (x_1^k - x_2^k)^T (x_1^k - x_2^k). \tag{1}$$

The conventional Euclidean distance assumes that different dimensions of x^i and x^j are perpendicular. This assumption does not hold for the vectorized images. This means that the Euclidean distance may not be suitable for the measurement of the distance/dissimilarity between two images. Since the Euclidean distance discards the image structures, it is unable to reflect the real distance between images [9]. Alternatively IMED [9] considers the angles between different dimensions by introducing the metric matrix G. The IMED $d_{IMED}^2(x_1, x_2)$ between images x_1 and x_2 is defined as:

$$\begin{aligned} d_{IMED}^2(x_1, x_2) &= \sum_{i=1}^{MN} \sum_{j=1}^{MN} g_{ij} (x_1^i - x_2^i)(x_1^j - x_2^j) \\ &= (x_1 - x_2)^T G (x_1 - x_2) \end{aligned} \tag{2}$$

where G is the metric matrix and g_{ij} is the metric coefficient specifying the spatial relationship between pixels p_i and p_j, x_1^i and x_2^i indicate the reference pixel and $x_1^j$and x_2^j indicate the neighboring pixels. The weight g_{ij} is defined as:

$$g_{ij} = f(d_{ij}^s) = \frac{1}{2\pi\sigma^2} \exp\left(\frac{(-d_{ij}^s)^2}{2\sigma^2}\right) \tag{3}$$

where d_{ij}^s is the spatial distance between the pixels p_i and p_j on the image and σ is the width parameter. For example, if p_i is at location (k,l) and p_j is at location (k′,l′) then d_{ij}^s is given by:

$$d_{ij}^s = \sqrt{(k-k')^2 + (l-l')^2} \tag{4}$$

As each summation in Eq. 2 clearly has a computational complexity of $O(MN)$ in the number of pixels $M \times N$ in the image, the computation of the overall distance $d_{IMED}^2(x_1, x_2)$ has a computational complexity $O(M^2N^2)$.

As IMED takes into account spatial relationship between all the pixels, it is not sensitive to small spatial deformation [9].

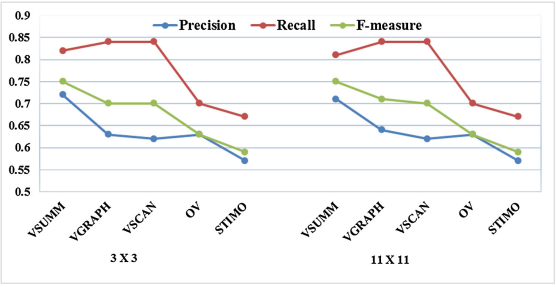

Fig. 1. Graphical representation of performance measures using different techniques and window sizes. Left: 3×3; Right: 11×11.

2.2 Efficient IMage Euclidean Distance (EIMED)

IMED [9] considers the spatial relationship between all the pixels and thus has an advantage that it can accommodate small deformation/movement of the objects in the images, at a high computational cost of $O(n^2)$ in the number of pixels in a given image. However, the movements of the objects in the neighboring frames in a video are usually small. On the other hand, Eq. 3 shows that the weight w_{ij} will exponentially decrease with regards to the distance d_{ij}. This implies that the distant pixels will make little contribution to the computation of $d_{IMED}^2(x_1, x_2)$. As a result, in this paper, we propose to consider only the neighboring pixels, just like a kernel with a size of $n \times n$ centred at the pixel of interest. If n increases, more neighboring pixels will be considered and the relative weights of the central pixels will decrease, and vice versa. This is proved in our experiments and will be discussed in Sect. 3 where we have identified that, 3×3 window size performs equally effective not only as 11×11 (see Fig. 1), but also achieved almost similar results as considering all the pixels within the images. The width parameter σ is set to 1 for simplicity. This way EIMED is computationally efficient in terms of extracting similar matching frames/images. The frame/image distance given in Eq. 2 is calculated for EIMED (3×3 window size) as depicted in Fig. 2 where red line indicates the reference pixel, blue lines indicate the neighboring pixels for that referenced pixel and the yellow square indicates the kernel size.

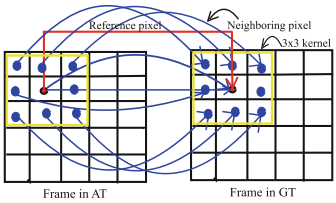

Fig. 2. Calculation of EIMED between two frames (Color figure online)

3 Experimental Results

In this section, we validate our proposed method for performance evaluation of video summaries using 50 videos selected from the Open Video Project[1]. The selected videos are in MPEG-1 format containing 30 fps with a resolution of 352×240 pixels. The videos include several genres (documentary, ephemeral, historical, lecture) and their duration varies from 1 to 4 min.

A comparative study was performed using five state-of-the-art techniques: VSUMM (Video SUMMarization) [3] based on color feature extraction and K-means clustering, VGRAPH [4] based on both color and texture features where key frames are extracted via clustering using K-Nearest Neighbor graph, VSCAN [5] based on modified Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN) utilizing both color and texture features, OV (Open Video Project) [12] based on a recursive multidimensional curve splitting algorithm, STIMO (STIll and MOving Video Storyboard) [6] based on color feature extraction and a fast clustering algorithm. The user study conducted by De Avila *et al.* [3] were used as ground truth summaries, where the user summaries were created by 50 users, each one dealing with 5 videos, meaning that each video has 5 different user summaries, so totally 250 summaries were created manually [3]. All the experiments were carried out on an Intel core i7, 3.60 GHz computer with 8 GB RAM. The performance metrics adopted in the proposed evaluation method are Fidelity, Precision, Recall and F-measure [10].

3.1 A Comparative Study

This section provides a comparative study of five state-of-the-art techniques: VSUMM [3], VGRAPH [4], VSCAN [5], OV [12], STIMO [6] using our proposed evaluation method. The experimental results in Table 1 show the mean performance measures achieved using various summarization techniques under our two-way evaluation method for different window sizes (3×3 and 11×11). It can be seen that VSUMM produced the best evaluation results since it eliminates meaningless and similar frames in the pre-processing and post-processing step respectively. The removal of meaningless frames in the pre-processing stage not only saves computation time but also improves the performance.

[1] Open Video Project. http://www.open-video.org.

Table 1. Mean performance measures achieved using various summarization techniques under our two-way evaluation method for different window sizes along with execution time t in seconds

Summarization techniques	# of videos	Window size	Mean				
			Fidelity	Precision	Recall	F-measure	t (s)
VSUMM	50	3×3	0.12	0.72	0.82	0.75	91
		11×11	0.11	0.71	0.81	0.75	260
VGRAPH	50	3×3	0.13	0.63	0.84	0.70	106
		11×11	0.12	0.64	0.84	0.71	293
VSCAN	50	3×3	0.12	0.62	0.84	0.70	120
		11×11	0.12	0.62	0.84	0.70	297
OV	50	3×3	0.11	0.63	0.70	0.63	85
		11×11	0.11	0.63	0.70	0.63	238
STIMO	50	3×3	0.12	0.57	0.67	0.59	87
		11×11	0.12	0.57	0.67	0.59	244

Even though VSUMM does not maintain temporal order as it employs K-means clustering for key frame extraction, we can conclude that from our evaluation results that VSUMM AT is very close to human perception. In contrast, STIMO lags behind, which may be improved by incorporating the elimination of meaningless frames during the pre-processing stage, though it removes possible redundancy during post-processing. In the case of VGRAPH, even though it eliminates the first frame of each shot as noise, it is worth incorporating the elimination of meaningless frames. With respect to VSCAN, using some other features like edge or motion instead of both color and texture may improve its performance. However, the key frames produced by OV are very concise which shows that some significant information might be missed leading to poor performance. It can be overcome by retrieving more key frames that well represent the entire video.

Figure 3 shows the automatic summaries obtained by different approaches (VSUMM, VGRAPH, VSCAN, OV, STIMO). It can be clearly seen that different techniques selected different numbers of frames and some of them are the same or similar, while the others are completely different or missing.

Figure 4 displays the user summaries for the same video, showing that even human users cannot agree completely on what frames should be selected as a summary of the entire video. This phenomenon shows that it is challenging to evaluate the keyframes selected by different techniques due to the fact that the ground truth is essentially missing or quite subjective.

Figure 5 shows VSUMM AT and its user summary #1 for the video *A New Horizon, segment 4* where it contains 13 AT frames and 6 GT frames, in which the green arrows show the 5 corresponding matches (such as region

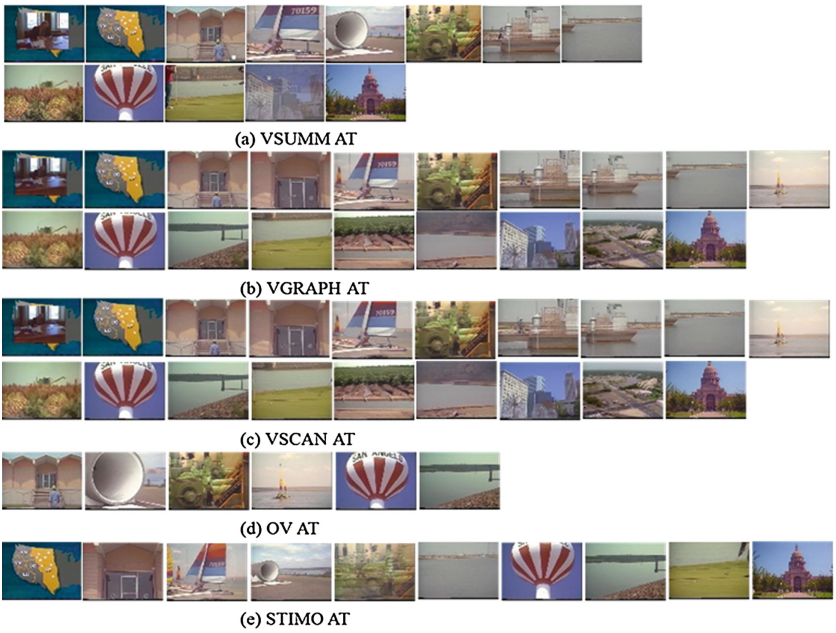

Fig. 3. Video summaries of various techniques for the video *A New Horizon, segment 4 (available at the Open Video Project)*

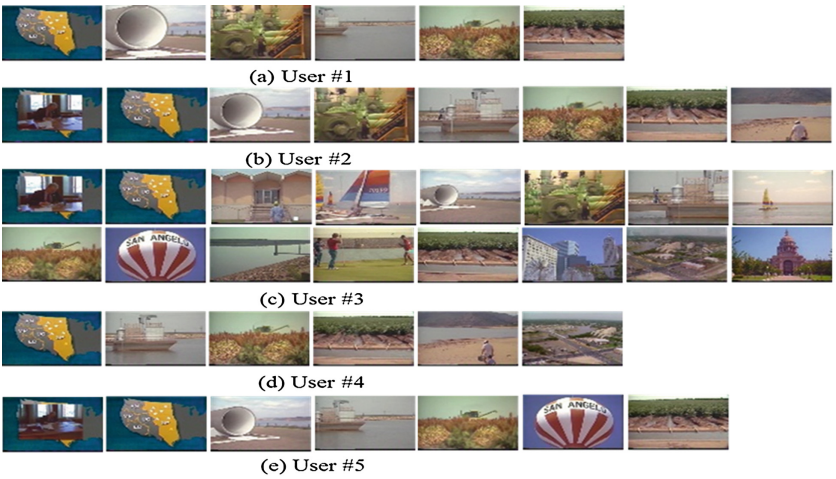

Fig. 4. User Summaries of the video *A New Horizon, segment 4 (available at the Open Video Project)*

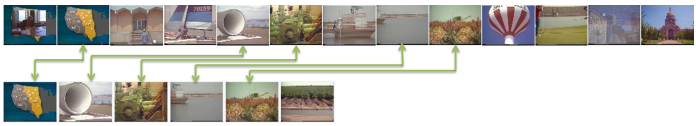

Fig. 5. VSUMM AT *(top)* and User summary #1 *(bottom)* of the video *A New Horizon, segment 4 (available at the Open Video Project)* (Color figure online)

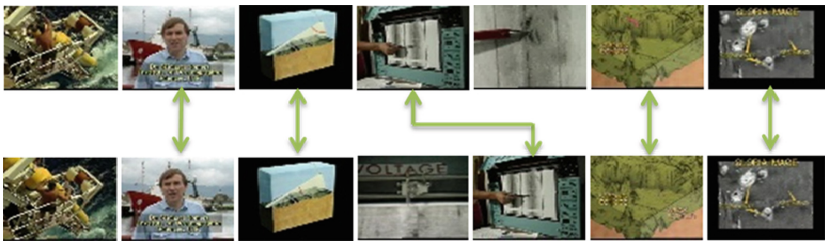

Fig. 6. VGRAPH AT *(top)* and User summary #5 *(bottom)* of the video *America's New Frontier, segment 4 (available at the Open Video Project)* (Color figure online)

map, pipeline, pumping plant, reservoir and agricultural land) between AT and GT. On the other hand, Fig. 6 shows VGRAPH AT and its user summary #5 for the video *America's New Frontier, segment 4* which contains 7 AT frames and 7 GT frames, in which the green arrows show the 5 corresponding matches (such as man with texts, sea floor geology, person pointing with pen, rocks & mountainscape geology and gloria image) between AT and GT. Even though the first frame of AT and GT in Fig. 6 appears to be similar at first sight to human eye, actually there is a slight variation of those frames, in the position of the man operating the ship. Our method detects successfully even this slight variation of position and considers those frames as distinct ones, rather than matched ones, thus providing reliable measurement of the performance of various video summary techniques.

To have an overall evaluation of the effectiveness of our method, we present the relative performance of different video summary techniques with some of the previous studies over the same dataset in Table 2. It can be seen that the mean F-measure of different techniques achieved by our proposed method is usually low, except for VSUMM. This means that the existing video summary techniques may not perform as well as expected. This is because our method discarded the false similarity matched frames between AT and GT, and thus provide more realistic evaluation of the performance of the video summary techniques. Such finding will be helpful for future researchers to investigate and develop more advanced techniques.

Table 2. The F-measure in percentage (%) of different video summary technique reported in the literature

Authors	# of videos	Mean F-measure				
		VSUMM	VGRAPH	VSCAN	OV	STIMO
Our method	50	75	70	70	63	59
Mahmoud [8]	50	72	75	77	67	65
Mahmoud *et al.* [4]	50	72	75	-	67	65

Table 3. The results of kernel size effect on the performance measurement of VSUMM AT against User Summary #5

Window size	# of videos	Mean		
		Precision	Recall	F-measure
3×3	50	0.71	0.83	0.75
11×11	50	0.71	0.82	0.75
$n \times n$	50	0.73	0.85	0.77

3.2 Computational Efficiency

From the quantitative comparison in Table 1 we can notice that the window sizes 3×3 and 11×11 perform almost equally effective in terms of accuracy but the average computational time for 3×3 window size was 1 min and 38 s whereas 11×11 window size took 4 min and 26 s, increasing computational time by 171%. On the other hand considering the spatial relationship of all the pixels for 50 videos, it took nearly 3 h for VSUMM AT with User summary #5. It achieves almost similar accuracy as 3×3 window size as shown in Table 3. Therefore to evaluate a single technique with all the 5 different user summaries, $n \times n$ window size would take nearly 15 h. This is almost intolerable. Thus we chose 3×3 window size as optimal, due to its accuracy and speed performance.

4 Conclusions

This paper has proposed a novel approach for the evaluation of automatic video summaries where the distance between the two frames are measured using EIMED. Due to the property of considering the spatial relationship between pixels, IMED is a preferred distance measure for images. EIMED considers only the neighboring pixels centered at the pixel of interest, rather than all the pixels, and thus gain computational efficiency. A comparative study based on a publicly accessible dataset shows that such distance did not sacrifice much in performance measurement, but gain significant computational efficiency. Based on the proposed method, our study showed that the existing techniques may not perform as well as expected, due to the crop up of false matched frames between

AT and GT. Furthermore, our study also produced a new ranking of the existing video summary techniques. Such findings will be useful for future researchers to develop more advanced techniques and carry out comparative studies among those different techniques.

Acknowledgements. The first author would like to thank for the award given by Aberystwyth University under the Departmental Overseas Scholarship (DOS) and partly funding by Object Matrix, Ltd on the project.

References

1. Pfeiffer, S., Lienhart, R., Fischer, S., Effelsberg, W.: Abstracting digital movies automatically. J. Vis. Commun. Image Represent. **7**, 345–353 (1996)
2. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **3**, 3 (2007)
3. De Avila, S.E.F., Lopes, A.P.B., da Luz, A., de Albuquerque Araújo, A.: VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recogn. Lett. **32**, 56–68 (2011)
4. Mahmoud, K., Ghanem, N., Ismail, M.: VGRAPH: an effective approach for generating static video summaries. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 811–818 (2013)
5. Mohamed, K.M., Ismail, M.A., Ghanem, N.M.: VSCAN: an enhanced video summarization using density-based spatial clustering. arXiv preprint arXiv:1405.0174 (2014)
6. Furini, M., Geraci, F., Montangero, M., Pellegrini, M.: STIMO: STILL and MOving video storyboard for the web scenario. Multimedia Tools Appl. **46**, 47–69 (2010)
7. Mei, S., Guan, G., Wang, Z., Wan, S., He, M., Feng, D.D.: Video summarization via minimum sparse reconstruction. Pattern Recogn. **48**, 522–533 (2015)
8. Mahmoud, K.M.: An enhanced method for evaluating automatic video summaries. arXiv preprint arXiv:1401.3590 (2014)
9. Wang, L., Zhang, Y., Feng, J.: On the euclidean distance of images. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1334–1339 (2005)
10. Kannappan, S., Liu, Y., Tiddeman, B.P.: A pertinent evaluation of automatic video summary. In: Proceedings of the 23rd International Conference on Pattern Recognition (2016, in press)
11. Li, J., Lu, B.L.: An adaptive image euclidean distance. Pattern Recogn. **42**, 349–357 (2009)
12. DeMenthon, D., Kobla, V., Doermann, D.: Video summarization by curve simplification. In: Proceedings of the Sixth ACM International Conference on Multimedia, pp. 211–218. ACM (1998)