VIDEO SUMMARIZATION BASED ON LOCAL FEATURES

Mohamed Massaoudi Sahbi Bahroun Ezzeddine Zagrouba

Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA) - LIMTIC Laboratory
Institut Supérieur d'Informatique (ISI), Université Tunis El-Manar
2 Rue Abou Rayhane Bayrouni, 2080 Ariana, Tunisia

med.massaoudi@yahoo.fr Sahbi.Bahroun@isi.rnu.tn ezzeddine.zagrouba@fsm.rnu.tn

ABSTRACT

Keyframe extraction process consists on presenting an abstract of the entire video with the most representative frames. It is one of the basic procedures relating to video retrieval and summary. This paper present a novel method for keyframe extraction based on SURF local features. First, we select a group of candidate frames from a video shot using a leap extraction technique. Then, SURF is used to detect and describe local features on the candidate frames. After that, we analyzed those features to eliminate near duplicate keyframes, helping to keep a compact set, using FLANN method. We developed a comparative study to evaluate our method with three state of the art approaches based on local features. The results show that our method overcomes those approaches.

Keywords

Video Summarization, Keyframe Extraction, Interest Points, SURF, FLANN.

1. INTRODUCTION

Videos have turned out to be the main source of information, learning and entertainment, with the growing advancements and progress in multimedia technologies. Daily, millions of videos are being uploaded on Internet consisting of news, sports clips, tutorials, lectures contents and many more. Content based retrieval of video has emerged as a growing challenge and therefore, automatic keyframes extraction; the main step for the efficient retrieval, video classification and story retrieval; has become so important and vital.

Using one keyframe as shot representation was considered by the majority of works found on the literature which defined, for example, as the shot first frame or median frame. Nevertheless just one frame, in most cases, is not able of fully representing the variety of information in a shot, usually composed by hundreds of images that can have different content [1].

Therefore, in this work, we propose a novel method for keyframe extraction based on local features. Due to their capabilities of retaining image semantics and providing robust descriptors, local features were the most reliable and widely applied method in the image retrieval field [2-5]. However, they have been poorly explored in the video keyframe extraction domain.

The rest of this paper is organized as follows. In Section 2 we briefly introduce three state of the art keyframe extraction approaches based on local features found in literature [6, 7, 8]. We will describes the proposed keyframe extraction method in section 3. Section 4 presents and analyzes the experimental results obtained. Finally, Section 5 concludes the paper and opens some perspectives of future work.

2. RELATED WORKS

In order to represent a shot, compact approaches are usually adopted which most of them are based on keyframes and color histograms. Those methods suffers a low representativeness and supposes that the video is already segmented into shots by a shot detection algorithm. Many works in the literature have been proposed for keyframes extraction based on local features. Furthermore, we discuss three keyframe extraction methods based on local features found in the literature.

Baber et al. [6] used SURF features to describe each extracted keyframe. First, the video is segmented into shots, then; the keyframes are defined as the shot median frame for each shot. Even if this approach has low computational cost, since it considers only a small fraction of the available frames, there is the issue of selecting an image, which is the median frame in this case, that does not represent the most relevant content of the shot.

Chergui et al. [7] consider that a relevant image contains rich visual details. Thus, they defined the keyframe as the frame with the highest number of points of interest in the shot. Despite using images content, it is not possible to guarantee that the frame with the highest number of points of interest is the most representative one in all cases. Besides, one image may not be enough to describe the diverse content

of some shots and important information can be lost. This method is also more computationally demanding, because the selection step involves processing all shot frames.

Tapu and Zaharia [8] extract a variable number of frames from each shot using a leap extraction technique. Then, each frame is compared with the existing keyframes already extracted. If the visual dissimilarity between them is significant, the frame is added to the keyframes set. After that, the extracted keyframes are described by SIFT. This method may have the advantage that not all shot frames are processed, but many parameters need to be set what can influence the quality of the shot representation.

The related work presented in this section show that the use of local features can be a substitute for keyframe representations. However, as discussed, the current approaches present problems of representativeness and computational costs leading to high processing times

3. PROPOSED APPROACH

We developed a keyframe extraction method based on local features, designed to deal with the problems identified in related work and discussed on at section 2, i.e., representativeness and computational cost due to high processing times.

3.1. Candidates Frames Selection

In order to select the best frames to be the keyframes of each shot, we, initially, select some frames into a Candidates Set (CS). The first frame to be included in the CS is defined as the shot first frame. This has the goal of guarantee that each shot will be represented by, at least, one keyframe. The next frames to be included in the CS follow a windowing rule. We defined a window of size n and the frames at positions n+1, 2n+1, 3n+1, and so on, are selected for later analysis. We set the fps value for n because within 1 second there is no significant variation on consecutive frames content.

Algorithm 1 Candidates frame generation

Require: Video V={$f_1, f_2, \ldots, f_n$}
1: fps = V.getFPS()
2: i = 1
3: while i < n do
4: cs.add(f_i)
5: i = i + fps
6: end while
7: return cs

3.2. Keyframes extraction

The next step is to extract SURF [9] features from the frames in the CS. The result is a number of feature vectors, of 64 dimensions, representing each frame. SURF features matching is faster compared to other descriptors such as SIFT. SURF features are also invariant to scale, rotation and partial illumination change [10]. The exact number of vectors varies according to the frames content but it is generally high. This is another reason to adopt the windowing rule (mentioned before) instead of to use all frames in the shot (see Figure 1).

Algorithm 2 Keyframes extraction

Require: Candidates Set $cs=\{cf_1, cf_2, \ldots, cf_m\}$
1: keyframes.add(cf_1)
2: i = 2
3: while i < m
4: U = extractSURF(cf_i)
5: isKeyFrame = True
6: for k = 1 to ks.size() do
7: V = extractSURF(cf_k)
8: M = matching(U,V)
9: y = $(1 - \frac{M}{|U|}) \times 100$
10: if y < 80% then
11: isKeyFrame = False
12: end if
13: end for
14: if isKeyFrame = True then
15: keyframes.add(cf_i)
16: end if
17: i = i + 1
18: end while
19: return keyframess

The first keyframe extracted is the first one in the CS. Then, each frame in the CS is analyzed according to the following criterion: it will be considered as a keyframe only if it has more than 80% (which is according to the literature the typical value used in vision applications) [11] of feature vectors different from each keyframe already extracted. The matching score y is defined as:

$$y = (1 - \frac{M}{|U|}) \times 100 \quad (1)$$

where M is number of matched features, |U| is number of features in the analyzed candidate frame. The reasoning behind this criterion is to avoid the extraction of similar keyframes, since similar keyframes do not add value to representativeness. We used the FLANN method proposed in [12] for automatically selecting the best matching method and its parameters for a given training set. It was shown to be fast in practice and is part of the OpenCV library.

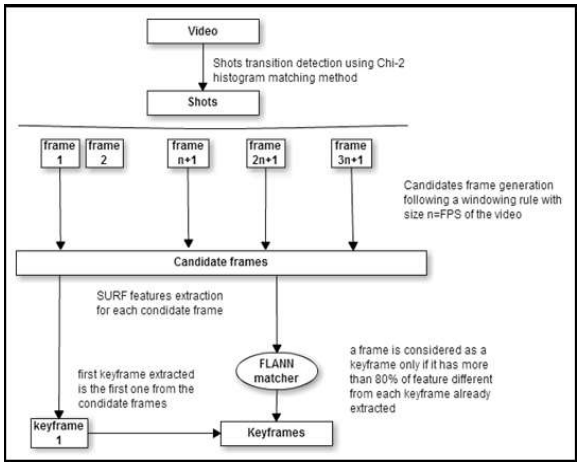

Figure 1. Proposed method steps.

3.3. FLANN matcher

FLANN stands for Fast Library for Approximate Nearest Neighbors which is a library of feature matching methods. It can provide automatic selection of index tree and parameter based on the user's optimization preference on a particular data-set. Automatic algorithm configuration allows a user to achieve high performance in approximate nearest neighbor matching by calling a single library routine. The user need only provide an example of the type of dataset that will be used and the desired precision, and may optionally specify the importance of minimizing memory or build time rather than just search time.

4. EXPERIMENTAL RESULTS

The efficiency of the proposed keyframe extraction method was evaluated by experimental tests on some videos (news, cartoons, games,…). These video present different challenges (camera motion, background-foreground similar appearance, dynamic background,…). Results proved that the method can extract efficiently keyframes resuming the salient semantic content of a video with no redundancy.

To verify the effectiveness of the proposed method, we first use qualitative evaluation since the subjective evaluation of the extracted key frame is efficient and it was used in many state of the art methods. In a second step, we will complete the evaluation with a quantitative study by calculating fidelity and compression rate. The use of quantitative and qualitative evaluation can prove the effectiveness of our proposed approach. The experiments were done on movies from YUV Video Sequences (http://trace.eas.asu.edu/yuv/) and some other standard test videos with different sizes and contents. In this paper we will show experiments done only on 7 movies as example. These movies were already segmented into shots by the $\chi 2$ histogram matching method [13]. Table 1 shows the number of frames and shots for the seven movies.

Movie	Frames	Shots
News	300	4
Bus	150	6
Foreman	297	3
Mother and Daughter	300	1
Suzie	150	4
Salesman	449	8
Carphone	382	7

Table 1. The videos characteristics

4.1. Validity Measures

For validity measures we used the fidelity and the compression rate.

4.1.1 Fidelity

The fidelity measure is based on semiHausdorff distance to compare each key frame in the summary with the other frames in the video sequence. Let $V_{seq}=\{F_1,F_2,\ldots F_N\}$ the frames of the input video sequence and let KF all keyframes extracted $KF=\{F_{K1}, F_{K2},\ldots, F_{KM}\ldots\}$. The distance between the set of key frames and F belonging to V_{seq} is defined as follows:

$$DIST(F,KF) = Min\{Diff(F,F_{Kj})\}_{,j=1\ to\ M} \quad (2)$$

Diff() is a suitable frame difference. This difference is calculated from their histograms: a combination of color histogram intersection and edge histogram based dissimilarity measure [14] .The distance between the set of key frames KF and the video sequence V_{seq} is defined as follows:

$$DIST(V_{seq},KF) = Max\{Diff(F_i,\mathrm{KF})\}_{,i=1\ to\ N} \quad (3)$$

So we can define the fidelity (FD) as follows:

$$FD(V_{seq},KF) = MaxDiff - DIST(V_{seq},KF) \quad (4)$$

MaxDiff is the largest value that can take the difference between two frames Diff (). High Fidelity values indicate that the result of extracted keyframes from the video sequence provides good and global description of the visual content of the sequence.

4.1.2 Compression Rate

Keyframe extraction result should not contain many key frames in order to avoid redundancy. That's why we should evaluate the compactness of the summary. The compression ratio is computed by dividing the number of key frames in the summary by the length of video sequence. For a given video sequence, the compression rate is computed as follows:

$$CR = 1 - \frac{card\{keyframes\}}{card\{frames\}} \quad (5)$$

Where card{keyframes} is the number of extracted key frames from the video. Card{frames} is the number of frames is the video.

4.2. Qualitative Evaluation

Now, we will present some results for 2 examples of videos. The first one is "news.mpg" which has 300 frames segmented into 4 shots. The figure 3 shows the 2 resulting key frames. As we can see the first image in figure 3 is the keyframe relative to the first shot of "news" video presented in figure 2 which is very logic, furthermore, the redundancy was eliminated.

The second video is "foreman.mpg" it is composed of 297 frames and segmented into 3 shots. The figure 5 shows the resulting keyframes for all the video. In the same way the first image of figure 5 is the keyframe relative to the first shot of "foreman" video (figure 4). Table 2 summarizes the number of key frames extracted for each video.

Movie	Number Of keyframes
News	2
Bus	4
Foreman	3
Mother and Daughter	1
Suzie	2
Salesman	3
Carphone	3

Table 2. Number of keyframes extracted per video

4.3. Quantitative Evaluation

We measured now for each movie, the fidelity and the compression rate (CR %). The table 3 illustrates these results.

Movie	Fidelity	CR%
News	0.80	99.33
Bus	0.69	97.33
Foreman	0.74	98.90
Mother and Daughter	0.77	99.60
Suzie	0.81	98.60
Salesman	0.77	99.32
Carphone	0.80	99.21

Table 3. Results in term of fidelity and compression rate

While looking to the results in Table 3 by the compression ratio (CR) values, it is clear that the proposed method minimizes considerably the redundancy of the extracted keyframes which guarantees encouraging compression ratios while maintaining minimum requirements of memory space. The Fidelity values confirm the same interpretation that we get by looking to the compression rate.

In order to give an objective evaluation, we compared the resulting quality measures of compression rate of our proposed method with three state of the art methods [6, 7, 8] and this for the seven tested videos in Table 3.

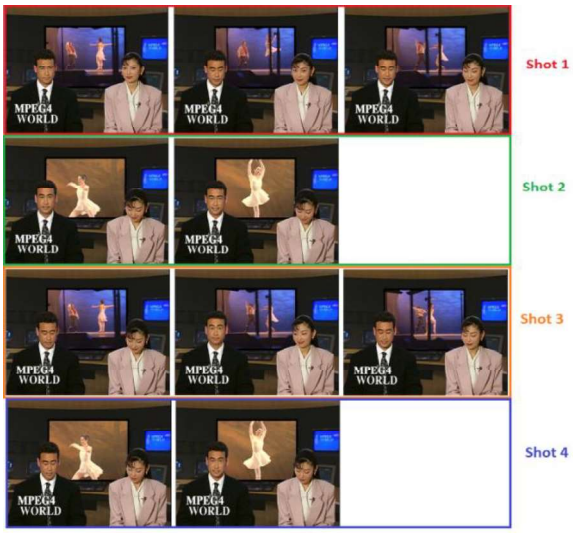

Figure 2. Shots of the video "news.mpg".

Figure 3. Keyframes of the video "news.mpg".

Figure 4. Shots of the video "foreman.mpg".

Figure 5. Keyframes of the video "foreman.mpg".

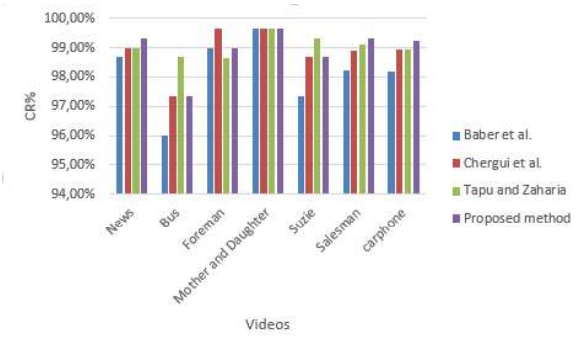

Figure 6: Comparison of the quality of the extracted keyframes in term of compression rate (CR%).

In Figure 6, we show a comparison between our proposed method and three state of the art methods in terms of compression rate. As the CR value is high as we have different keyframes. We can see in Figure 6 that our proposed method reduced considerably the redundancy of extracted keyframes.

5. CONCLUSIONS

In this paper, we have proposed a simple and effective technique for keyframe extraction based on SURF local features and using the FLANN matching method. Firstly, candidate frames are selected adaptively using a leap extraction method. Each candidate frame is described by SURF local features vectors. Secondly, keyframes for each shot are selected from the candidate frames set using FLANN method to discard any duplicated keyframes. The proposed approach proved to have superior effectiveness to three state of the art related work, i.e., gives a set of image that covers all significant events in the video while minimizing information redundancy in keyframes.

As a perspective, we consider developing a complete system for still image-based face based on visual summary which is composed by faces from the extracted keyframes. The user can initiate his visual query by selecting one face and the system respond with videos which contains that face.

6. REFERENCES

[1] Souza, T.T. and Goularte, R. 2013. Video Shot Representation Based on Histograms. Proceedings of the 28th ACM Symposium on Applied Computing (Coimbra, Portugal, 2013), 961–966.

[2] Baber, J., Satoh, S., Afzulpurkar, N. and Keatmanee, C. 2013. Bag of Visual Words Model for Videos Segmentation into Scenes. Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service (New York, NY, USA, 2013), 191–194.

[3] Blanken, H.M., Vries, A.P., Blok, H.E. and Feng, L. 2010. Multimedia Retrieval. Springer.

[4] Lowe, D.G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. 60, 2 (2004), 91–110.

[5] Lowe, D.G. 1999. Object recognition from local scale-invariant features. Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on (Kerkyra, Greece, 1999), 1150 – 1157 vol.2.

[6] Baber, J., Afzulpurkar, N. and Bakhtyar, M. 2011. Video segmentation into scenes using entropy and SURF. Emerging Technologies (ICET), 2011 7th International Conference on (2011), 1–6.

[7] Chergui, A., Bekkhoucha, A. and Sabbar, W. 2012. Video scene segmentation using the shot transition detection by local characterization of the points of interest. Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on (2012), 404–411.

[8] Tapu, R. and Zaharia, T. 2011. A complete framework for temporal video segmentation. Consumer Electronics – Berlin (ICCE-Berlin), 2011 IEEE International Conference on (2011), 156–160.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," Comput. Vis. Image Underst., vol. 110, no. 3, pp. 346– 359, 2008.

[10] S. Bahroun, H. Gharbi and E. Zagrouba, "Local query on satellite images based on interest points," 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, 2014, pp. 4508-4511.

[11] Dror Aiger, Efi Kokiopoulou, Ehud Rivlin. Random Grids: Fast Approximate Nearest Neighbors and Range Searching for Image Search. The IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3471-3478

[12] M. Muja, D. G. Lowe, Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration, in International Conference on Computer Vision Theory and Applications (VISAPP'09), 2009

[13] Cai and al, 2005. A Study of Video Scenes Clustering Based on Shot Key Frames . Series Core Journal Of Wuhan University (English) Wuhan University Journal Of Natural Sciences Pages 966-970

[14] Ciocca, G., and Schettini, R. 2006. An innovative algorithm for key frame extraction in video summarization. J. of Real-Time Image Processing 1(1): 69-88.