KEY FRAMES EXTRACTION USING GRAPH MODULARITY CLUSTERING FOR EFFICIENT VIDEO SUMMARIZATION

Hana Gharbi, Sahbi Bahroun, Mohamed Massaoudi and Ezzeddine Zagrouba

Laboratoire LIMTIC, Institut Supérieur d'Informatqiue, Université de Tunis El Manar.
2 Rue Abou Rayhane Bayrouni, 2080 Ariana, Tunisie.

ABSTRACT

Keyframe extraction is one of the basic procedures relating to video retrieval and summary. It consists on presenting an abstract of the video with the most representative frames. This paper presents an efficient keyframe extraction approach based on local description and graph modularity clustering. The first step is to generate a set of candidate keyframes using a windowing rule in order to reduce the data to be examined. After that, detect interest points in these set of images. Then compute repeatability between each two images belonging to the candidate set and stocks these values in a matrix that we called repeatability matrix. Finally, the repeatability matrix is modelled by an oriented graph and we will select keyframes using graph modularity clustering principle. The experiments showed that this method succeeds in extracting keyframes while preserving the salient content of the video. Further, we found good values in term of precision, PSNR and compression rate.

Index Terms— key frame extraction, interest points, local features, repeatability

1. INTRODUCTION

Video applications, which are greatly growing, have encouraged the increasing tools for efficient summarization, indexing and retrieval of video data. Keyframe extraction is an important step in video retrieval and summary since it can provide a concise and accurate representation of the original video. This process reduces significantly the amount of data that must be examined and which is required in many types of applications depending to the needs of the user. Key-frame video abstraction consists of converting an entire video to a few numbers of representative images. These images should maintain the salient content of the video while eliminating all redundancy in the content. Most of the work in video retrieval and summarization extracts key frames for each shot. A Shot is composed by a sequence of frames. A video shot is defined as a sequence of frames captured by one camera in a single continuous action in time and space [1]. Generally, it is a group of frames that have consistent visual characteristics, such as colour, texture, and motion.

This paper treats the problem of video key frame extraction. The main goal of the process is minimizing information redundancy by finding the minimal set of image that cover all significant events in the video. Although a number of techniques have been found in the literature for key frame extraction, the majority of successful ones are computationally expensive [1] and few of them give importance to local description. They are in majority based on global image description extracted by computing similarity measure between descriptors. We propose in this paper a simple and effective technique for key frame extraction based on local features description, interest points matching and graph modularity clustering.

In Section 2, we present some recent approaches of key frame extraction for video summary and retrieval. In Section 3 we describe the key frame extraction method proposed. The results and observations of the new key frame extraction method comparing with other important works in the same field are discussed in Section 4. We conclude in Section 5.

2. RELATED WORKS

While we are faced to a huge volume of video content, video summarization plays an important role in efficient storage, and retrieval without losing important resources like man power, time and storage. The key frame extraction is an important technique for video summarization. We will present some novel key frame extraction methods: X. Liu, et al [2] proposed a method based on MAP: Maximum a Posteriori in order to estimate the positions of key frames. N. Ejaz, et al. [3] proposed an aggregation mechanism in order to combine the visual features extracted from the correlation of channels of RGB color, the moments of inertia and the color histogram to extract key frames. Q. Xu, et al [4,5] developed a Jensen_Rényi divergence, Jensen_Shannon divergence and Jensen_Tsallis divergence-based approach to measure the difference between extract key frames and neighboring frames. J.L. Lai, et al [6] used a saliency-based visual attention model and selected as key frames the frames with maximum saliency value. M. Kumar,

et al in [7] analyzed the spatio_temporal information of the video using sparse representation and used a normalized clustering method to generate clusters; then the middle frame in each temporal order-sorted cluster was chosen as a key frame. In [8] a graph connectivity technique and a dominant set clustering method were combined for automatic keyframes selection. Shao et al. [9] proposed a cluster-based keyframe selection using visual and textual features. In [10], Ngo et al analyzes video structure by graph modelling then the video summary is generated according to this structure and the motion attention values for video shots.

Despite various methods existing in literature, the problem of key frame extraction remains a challenging due to the complexity and diversity of video content. Local features can give an accurate solution for these problems but existing method using this alternative are not robust: Chergui et al. [11] adopted a strategy that select a single keyframe to represent each shot. They consider that the key frame contains richest visual details and thus it is the frame with the highest number of points of interest in the shot which is not true in all cases. Besides, one image is not always enough to describe the diverse content of some shots and important information can be lost. Also it is more computationally demanding, because the selection step involves processing all shot frames. Tapu et al [12] developed an approach to extract a variable number of keyframes from each shot. Using N as window size parameter, the first frame is extracted after N frames from the detected shot transition. Next, they analyze frames located at integer multipliers of the window of size N. These images are compared with the existing keyframes set which are already extracted. Then if the visual dissimilarity between them is significant, the current image is added to the keyframes set. Then, they eliminate irrelevant frames, computing interest points using SIFT descriptor. If the keypoints number is equal to zero, the image is discarded. Then, the keyframes are described using SIFT features. This algorithm of keyframe extraction has the advantage that not all shot frames are processed. Yet, many parameters need to be set, what can influence the quality of the shot representation. Gharbi et al. proposed in [13] an approach which is based on interest points description and repeatability measurement. Before key frame extraction, the video is segmented into shots. Then, for each shot, detect interest points in all images. After that, calculate repeatability matrix for each shot. Finally, apply PCA and HAC to extract key frames.

After this study of the related work of key frame extraction, we can see that using local features can be good alternative for keyframe representations. However, as discussed, the current methods present problems of representativeness and sometimes computational costs which can lead to high processing times and data dimension [11],[12]. For exemple the work by chergui et al [11] is computationally demanding since the selection step involves processing all shot frames, we involve this step by using the windowing rules and treat only the candidate set resulting. The work by Gharbi et al [13] suffers from the problem of the loss of information by using PCA and redundancy since it treats separately shots, so each shot will have necessary at least one keyframe. The work presented here avoid these problems by reducing the table dimension using the windowing rules. Also it represents the table by an oriented graph which gives a good agreement between local features and complexity. These advantages will be proved in experimental results.

3. PROPOSED APPROACH

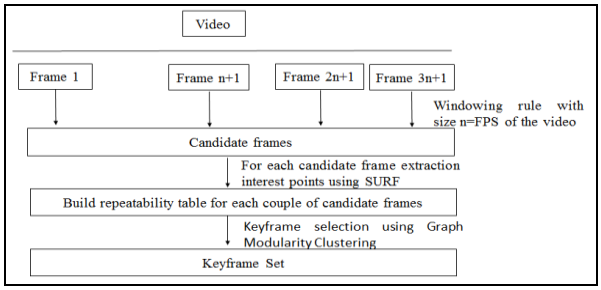

Fig. 1. Proposed approach

3.1. Candidates frames generation

In order to select the best frames to be the keyframes, we, initially, select some frames into a Candidates Set (CS). The first frame to be included in the CS is defined as the first video frame. Then the next frames to be included in the CS follow a windowing rule. We defined a window of size n and the other frames at positions: n+1, 2n+1, 3n+1, and so on, are selected for later analysis. We set the fps (frame per seconds) value for n because within 1 second there is no significant variation on consecutive frames content.

3.2. SURF detector

The next step is to extract SURF [14] 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 [15]. 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 in 3.1 instead of to use all frames in the shot.

3.3. Build the repeatability table

After detecting interest points in each frame of (CS) in the video shots, we will compute the repeatability matrix. Repeatability is a criterion which proves the stability of the interest point detector. It is the average number of corresponding interest points detected in images under noise

or changes undergone by the image [16]. This matrix is built from all images belonging to (CS). We must compute repeatability between each two part of the (CS) frames.

Algorithm 1:

Inputs: M: matrix with N x N dimension
N: number of (CS) in the video
Outputs: M: matrix filled with the repeatability values

```
1:  Begin
2:      M[i][j]= M[N][ N]
3:      for (int i = 0; i < N ; i++)
4:      for (int i = 0; i < N ; i++)
5:       // apply matching algorithm for this two images
6:       //compute the repeatability between I and J frames
7:       M[i][j]=Repeatability i,j
8:       End
9:       End
10: End
```

Our goal now is to detect the keyframes from this repeatability table and in order to reduce time and complexity we will resort to model this table into an oriented graph.

3.4. Keyframe selection using Graph Modularity Clustering

In this part, we will consider the repeatability matrix as an adjacency matrix and model it by an oriented graph. This graph is called the video similarity graph (VSG). We built it using images from the candidate frame set as its vertices. It is represented by G = (V, E, W), where V: the set of nodes, E: the set of edges connecting the nodes and W: the set of weights corresponding to the strength of edges. The weights W_{ij} between two frames is defined as the value from the repeatability table between candidate frames i and j. In the VSG graph, edges can be grouped into intra-cluster edges (edges whose end points are at the same cluster) and inter-cluster edges (edges whose end points are at different clusters). The objective is to preserve the intra-cluster edges and remove the inter-cluster ones. This will connect the individual clusters in an efficient manner. The principle is to prune certain edges depending on the difference between edge weights, until there is no improvement in graph modularity [17]. Modularity M (c_1, c_2, · · · , c_k) of a graph clustering over k known clusters c_1, c_2, · · · , c_k is defined as:

$$M(c_1, c_2, ..., c_K) = \sum_{i=1}^{k} \delta_{i,i} - \sum_{i=j} \delta_{i,j} \qquad (1)$$

Where $\delta_{i,j} = \sum_{\{u,v\} \in E, v \in c_i, u \in c_j} w(v,u)$, with each edge $\{v,u\} \in E$, E included at most once in the computation. High value of modularity indicates good clustering. Remaining connected components of the final VSG after end of edge pruning represent individual clusters. Algorithm 2 summarizes the steps involved in graph modularity clustering.

Algorithm 2: Graph Modularity Clustering

Inputs: Video Similarity Graph (VSG),E,W
M[N][N]; //repeatability table
N; // number of (CS) in the shot;
OUTPUT: Clusters{c_1, c_2, · · · , c_k}

```
1:  for i = 1 to N
2:  for j = 1 to N
3:  Dev_ij ≜ 1− M[i][j]
4:  End
5:  End
6:  Repeat
7:  Select edges which has high value of Dev then
     remove the edge from VSG
8:  Find connected components from the VSG
9:  Calculate Modularity (M)
10: Until no improvement in Modularity over two
     successive iterations.
11: Obtain individual clusters from final VSG
12: End
```

The frames, which are closest to the centroids of each cluster, are deemed as keyframes. Finally, the key frames are arranged in a temporal order to make the produced summary more understandable.

4. EXPERIMENTAL RESULTS

To evaluate the efficiency of our proposed key frame extraction method, we did experimental tests on some videos (news, cartoons, games,…). These video illustrate different challenges (camera motion, background-foreground similar appearance, dynamic background,…). To verify the robustness of the key frame extraction proposed method, 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. Results prove that the method is able to extract efficiently few key frames resuming the salient semantic content of a video (example: Fig. 2and 3). Notice that in the case of a reduced number of key-objects within the input video, our method extracts only key frames which are relevant and non-redundant.

Fig. 2. Key frame extraction by the proposed method from the standard video "foreman.mpg"

Fig.3. Key frames extracted from the video "filinstone.mpg"

For an objective assessment, we used the signal to noise ratio (PSNR). In fact, for each couple (F_u, F_v) of selected key frames (of size N M), we measure the PSNR between them (2) and the mean value is recorded for each studied video (Fig. 4).

$$PSNR(F_u, F_v) = 10.\log\left(\frac{N.M.255^2}{\sum_{x=1}^{N}\sum_{y=1}^{M}(F_u(x,y) - F_v(x,y))^2}\right) \quad (2)$$

The more key frames F_u and F_v are similar, the more PSNR value is high. Infinite values reflect the redundancy of the extracted key frames and reduced PSNR values indicate the diversity of these key frames. The recorded PSNR values by the proposed method are minimal compared to the other methods (Fig. 4). These values confirm that our method (PA) extracts the most significant and relevant key frames while minimizing redundancy. We used also the compression ratio (CR):

$$CR = 1 - \frac{card\{(Keyframes)\}}{card\{(frames)\}} \quad (3)$$

From (fig 5) it is clear that the proposed method (PA) minimizes considerably the redundancy of the extracted key frames, what guarantees encouraging compression ratios while maintaining minimum requirements in terms of memory space.

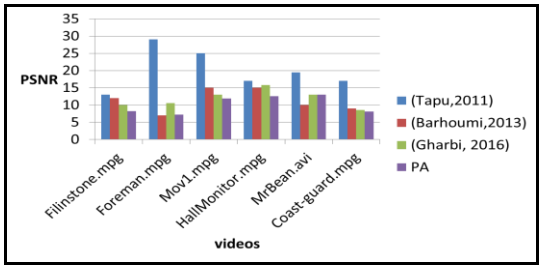

Fig. 4. Comparison of the quality of the produced results in term of PSNR values

We calculate also the precision average of the proposed approach (>85%): it is always higher than those of the compared ones (which varies between 69% and 81%).

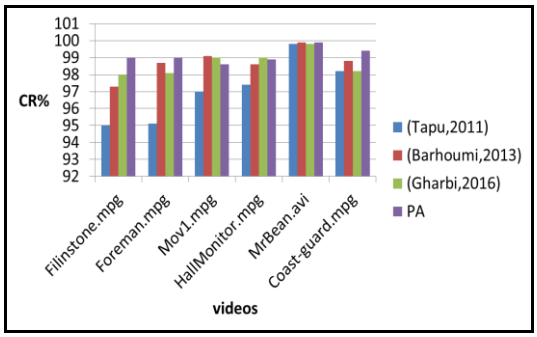

Fig. 5. Comparison of the quality of the produced results in term of compression rate (CR) values

All these results demonstrate the feasibility and efficiency of the proposed method. Our method can offer us a video summary with a no redundant key frames since our approach is based on oriented graphs. All similar images will be presented by one key frame. Also, our approach is with low computational cost since it is based on Graph Modularity Clustering.

5. CONCLUSION

In this paper, we presented a new approach for key frame extraction based on local image description "interest points", repeatabily matrix and modularity . The experiments showed that the proposed algorithm gives a set of images that cover all significant events in the video while minimizing information redundancy in these key frames by introducing Graph Modularity Clustering method. Although a number of techniques have been studied in the literature for key frame extraction, most of them are computationally expensive [4] and few of them gives important to local description. The results prove that local description can be good alternative in keyframe extraction field.

As a perspective of our work, after extracting key frames from all the videos in the database, we will try to give to the user the zero page of the video database. This zero page will contain the visual summary which will be composed by the most representative objects in the videos database. The user can initiate his visual query by selecting one or some of these objects by composing a mental query image.

5. REFERENCES

[1] M. Furini, F. Geraci, and M., Pellegrini, "M., STIMO: STIll and moving video storyboard for the web scenario," *Multimedia Tools and Applications*, pp 47–69, 2010.

[2] X. Liu, M. L Song, L. M. Zhang and S. L. Wang, "Joint Shot Boundary Detection and Key Frame Extraction", *Proceedings of the 21st International Conference on Pattern Recognition*, pp. 2565-2568, 2012.

[3] N. Ejaz, T. B. Tariq and S. W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism", *Journal of Vision Communication and Image Representation*, vol. 23, pp. 1031-1040, 2012.

[4] Q. Xu, Y. Liu, X. Li, Z. Yang, J. Wang, M. Sbert and R. Scopigno, "Browsing and exploration of video sequences: A new scheme for key frame extraction and 3D visualization using entropy based Jensen divergence", *Information Sciences*, vol. 278, pp. 736-756, 2014.

[5] Q. Xu, Y. Liu, X. Li, Z. Yang, J. Wang, M. Sbert and J. F. Li, "Key frame selection based on JensenRényi divergence", *Proceedings of the 21st International Conference on Pattern Recognition* (ICPR), pp. 1892-1895, 2012.

[6] J. L. Lai and Y. Yi," Key frame extraction based on visual attention model", *Journal of Vision Communication and Image Representation*, vol. 23, pp. 114-125, 2012.

[7] M. Kumar and A. C. Loui, "Key Frame Extraction from Consumer Videos Using Sparse Representation", *Proceedings of the 18th IEEE International Conference on Image Processing* (ICIP), pp. 2437-2440, 2011.

[8] D. Besiris, A. Makedonas, G. Economou, and S. Fotopoulos, "Combining graph connectivity & dominant set clustering for video summarization," Multimedia Tools and Applications, vol. 44, pp. 161–186, 2009

[9] J. Shao, D. Jiang, M. Wang, H. Chen, and L. Yao. Multi-video summarization using complex graph clustering and mining. In ComSIS, 2010

[10] C. W. Ngo, Y. F. Ma, and H. J. Zhang, "Automatic video summarization by graph modeling," in Proceedings of the 9th IEEE International Conference on Computer Vision, 2003, pp. 104–109.

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

[12] Tapu, R. and Zaharia, T. "A complete framework fortemporal video segmentation", *Consumer Electronics ICCE*, 156–1603, Berlin, 2011.

[13] H.Gharbi, S.Bahroun and E.Zagrouba, "A Novel Key Frame Extraction Approach For Video Summarization," *International Joint Conference on Computer Vision Theory and Applications*,pp 148-155, Rome 2016.

[14] 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.

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

[16] Schmid, C., Mohr, R., Bauckhage, C., " Evaluation of Interest Point Detectors", *International Journal of Computer Vision*, 2000.

[17] S.E .Schaeffer,"Graph clustering", *Computer Science Review* 1, pp 27–64 , 2007.