Key Frames Extraction Based on Local Features for Efficient Video Summarization

Hana Gharbi(✉), Mohamed Massaoudi, Sahbi Bahroun, and Ezzeddine Zagrouba

Research Team Systèmes Intelligents en Imagerie et Vision Artificielle
SIIVA– RIADI Laboratory ISI,
2 Rue Abou Rayhane Bayrouni, 2080 Ariana, Tunisia
hanagharbi@yahoo.fr, Sahbi.Bahroun@isi.rnu.tn,
ezzeddine.zagrouba@fsm.rnu.tn

Abstract. Key frames are the most representative images of a video. They are used in different areas in video processing, such as indexing, retrieval and summarization. In this paper we propose a novel approach for key frames extraction based on local feature description. This approach will be used to summarize the salient visual content of videos. First, we start by generating a set of candidate keyframes. Then we detect interest points for all these candidate frames. After that we will compute repeatability between them and stock the repeatability values in a matrix. Finally we will model repeatability table by an oriented graph and the selection of keframe is inspired from shortest path algorithm A*. Realized experiments on challenging videos show the efficiency of the proposed method: it demonstrates that it is able to prevent the redundancy of the extracted key frames and maintain minimum requirements in terms of memory space.

Keywords: Key frame extraction · Interest points · Local features · Repeatability

1 Introduction

Compared with text, audio and image, videos become the main source of information due to its abundant amount of information and intuitive experience. With rapid progress of computer and network technologies, millions of videos are daily being uploaded on Internet consisting of news, tutorials, sports clips, lectures contents and many others. 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.

A successful video summarization process aims to obtain a compact representation, which should be used to properly characterize videos. Key frames are a subset of still frames extracted from different video shots, and can be theoretically defined as the most representative and informative frames that maintain the salient content of the video. According to the definition, the purpose of key frame extraction algorithm is to extract correct and proper key frames from each video, which can perfectly represent the whole visual contents of the shot while eliminating all redundancy.

J. Blanc-Talon et al. (Eds.): ACIVS 2016, LNCS 10016, pp. 275–285, 2016.
DOI: 10.1007/978-3-319-48680-2_25

In this paper, we propose a novel keyframe extraction method based on local features. Local features have being applied successfully in the image retrieval domain, mainly due their capabilities of providing robust descriptors against different transformation types (rotation, viewpoint changes,…) [1–3]. However, in spite of their importance, local features have been poorly explored in the video keyframe extraction field.

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

2 Related Works

While we are faced to a huge volume of video content, video summarization plays an important role in efficient storage, quick browsing, and retrieval of large collection of video data without losing important resources like time, man power and storage [4]. The video can be seen as a combination of frames that is called GOP (groups of pictures). The key frame extraction is an important technique for video summarization. We can summarize the traditional key frame extraction for video summarization methods in five categories:

Shot boundary based methods [5, 6] suppose that the semantic and visual contents in one shot are mainly stable and change softly, only three frames which are: the first, the last and the middle frames in each shot are selected as the key frames, which is certainly not robust to majority of videos.

Motion analysis-based methods [7–9]: key frames are selected in local minimum through the computation of optical-flow. However, the computational cost is huge and the results aren't always accurate.

Clustering based methods and visual content based methods [10, 11]: they use the difference between adjacent frames to select key frames, and these two methods can be easily affected by noise and motion.

Event/object based methods [12, 13]: These algorithms jointly consider key-frame extraction and object/event detection in order to ensure that the extracted key-frames contain information about objects or events. Calic and Thomas [14] use the positions of regions obtained using frame segmentation to extract key-frames where objects merge. The merit of the object/event-based algorithms is that the extracted key-frames are semantically important, reflecting objects or the motion patterns of objects. The limitation of these algorithms is that object/event detection strongly relies on heuristic rules specified according to the application.

Trajectory based methods [15]: These algorithms represent each frame in a shot as a point in the feature space. The points are linked in the sequential order to form a trajectory curve and then searched to find a set of points which best represent the shape of the curve. Calic et al. [15] generate the frame difference metrics by analyzing statistics of the macro block features extracted from the MPEG compressed stream. The merit of the curve simplification-based algorithms is that the sequential information is kept during the key-frame extraction. Their limitation is that optimization of the best representation of the curve has a high computational complexity.

In this work, we will focus in particular in feature based methods to extract key frames. We will try to present some novel key frame extraction methods based on local description. Liu, et al. [16] proposed a method based on Maximum a Posteriori (MAP) to estimate the positions of key frames. Ejaz, et al. [17] proposed an aggregation mechanism to combine the visual features extracted from the correlation of RGB color channels, the color histogram and the moments of inertia to extract key frames. Xu, et al. [18] developed a Jensen–Shannon divergence, Jensen–Rényi divergence and Jensen–Tsallis divergence-based approach to measure the difference between neighboring frames and extract key frames. Lai, et al. [19] used a saliency-based visual attention model and selected the frames with maximum saliency value as key frames. Kumar, et al. [20] analyzed the spatio-temporal information of the video by sparse representation and used a normalized clustering method to generate clusters; the middle frame in each temporal order-sorted cluster was selected as a key frame. Sargent et al. [28] proposes a novel scalable summary generation approach based on the On-Line Analytical Processing data cube. Such a structure integrates tools like the drill down operation allowing to browse efficiently multiple descriptions of a dataset according to increased levels of detail.

Despite that many methods have been presented in the literature, key frame extraction remains a challenging and difficult problem due to the complexity and diversity of video content. Most of the feature extraction techniques are based on global feature extraction from each shot in the key frame. Local features can give an accurate solution for these problems. Chergui et al. [21] adopted a strategy that select a single keyframe to represent each shot. But this key frame extraction method is not robust. They consider that the key frame is the relevant image that contains richest visual details. Thus, they defined the key frame 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 et al. [22] developed an approach to extract a variable number of keyframes from each shot. Using a window size parameter N, the first frame is selected N frames after a detected shot transition. Next, they analyze images located at integer multipliers of the window size N. These images are compared with the existing keyframes set already extracted. If the visual dissimilarity (defined as the chi-square distance of HSV color histograms) between them is significant (above a pre-established threshold), the current image is added to the keyframes set. Then, they discard irrelevant frames, computing points of interest with SIFT descriptor. If the number of keypoints is zero, the image is removed. After that, the keyframes are described by SIFT features. This keyframe extraction method has the advantage that not all shot frames are processed. However, many parameters need to be set (window size N, dissimilarity threshold, histograms quantization), what can influence the quality of the shot representation. Gharbi et al. proposed [25] 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.

This approach shows good results in comparison with state of the art methods but it suffers from some problems like the loss of information by using PCA and redundancy since it treats separately shots, so each shot will have necessary at least one keyframe.

After this study of the related work of key frame extraction, we can see that different methods are either too naïve or too complex. The most simple of these techniques sorely compromise the key frames extracted quality and the most sophisticated ones are computationally very expensive. Also, some of these methods give us key frames with approximately the same content.

As we can see also, the related work using local features can be good alternative for keyframe representations. However, as discussed, the current approaches present problems of representativeness and, sometimes, computational costs leading to high processing times. Our proposed work gives a good agreement between local features, quality and complexity of results and this will be proved in experimental results.

3 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 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 s there is no significant variation in consecutive frames content.

3.2 SURF Detector

The next step is to extract SURF [23] features from the frames in the CS. The result is a number of feature vectors representing each frame, each one is of 64 dimensions. SURF features matching is faster compared to other descriptors such as SIFT [3]. 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 Sect. 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 using the SURF matching results. 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 [24]. This matrix is built from all images belonging to (CS). We must compute repeatability between each two part of the (CS) frames.

```
Inputs:
  M: matrix with N x N dimension
  N: number of (CS) in the video
Outputs:
  M: matrix filled with the repeatability values
Begin
M[i][j]= M[N][ N]
for (int i = 0; i < N ; i++)
  for (int i = 0; i < N ; i++)
  // apply matching algorithm for the two images
  // compute the repeatability between I and J frames
  M[i][j]=Repeatability i,j
  End
End
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 Shortest Path Algorithm

In this part we will consider the repeatability table as an adjacency matrix and we will model it by oriented graph: Frames are considered as vertices and the edges are weighted by repeatability values. Our method is inspired from shortest path algorithm A* which is a best-first search, meaning that it solves problems by searching among all possible paths to the goal for the one that incurs the smallest cost. In our case the smallest cost is the repeatability values. The path is directed and requires that consecutive vertices be connected by an appropriate directed edge since repeatability values are ordered in chronological sense. After founding the first occurrence of the minimum value (min-value) of the repeatability table we do the algorithm explained below:

```
Inputs
M: matrix filled with the repeatability values
KS= Ø ;         // Keyframes set
Nb_Lign= Nb_Col=N; // number of (CS) in the shot;
i=j=0;
Outputs:
KS: Keyframes set
Begin
While  (j<N) do
 While (i<N) do
    If T[i][j]==minvalue
         Add j to KS;
         j=i;
    else i++;
    End if
 End while
 j++; i=j;
End while
End
```

The use of shortest path algorithm to select relevant key frames help us to eliminate redundancy and to enhance accuracy with best time cost and complexity.

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,…). Experimental results proved that our method can extract efficiently key frames resuming the salient content of a video with no redundancy.

To verify the robustness of the key frame extraction proposed method we use qualitative and quantitative evaluation of the extracted key frames in order to enhance the proof of the effectiveness of our proposed approach.

In experimental setup, 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. The shot detection is based on the $\chi 2$ histogram matching [27]. Table 1 shows the number of frames and shots for the 6 movies:

Table 1. The video characteristics

Movie	Nb frames	Nb shots
Filinstone.mpg	510	10
Foreman.avi	297	5
Mov1.mpg	377	6
HallMonitor.mpg	299	4
MrBean.avi	2377	8
Coast-guard.mpg	299	2

4.1 Validity Mesures

Fidelity. The fidelity measure is based on semi-Hausdorff 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 key frames extracted $KF = \{F_{K1}, F_{K2},\ldots, F_{KM}\ldots\}$. The distance between the set of key frames and F belonging toV_{seq} is defined as follows:

$$DIST(F, KF) = Min\left\{(Diff(F, F_{Kj})\right\},\ j = 1\,to\,M \tag{1}$$

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 [13]. The distance between the set of key frames KF and the video sequence V_{seq} is defined as follows:

$$DIST(V_{seq}, KF) = Max\{DIST(F_i, KF)\},\ i = 1, .., N \tag{2}$$

So we can define the fidelity as follows:

$$FIDELITY\ (V_{seq}, KF) = MaxDiff - DIST(V_{seq}, KF) \tag{3}$$

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

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 the video sequence. For a given video sequence, the compression rate is computed as follows:

$$CR = 1 - \frac{card\{(Keyframes)\}}{card\{(frames)\}} \tag{4}$$

Where card(keyframes) is the number of extracted key frames from the video. Card (frames) is the number of frames in the video

Signal to noise ratio. We calculate also the signal to noise ratio (PSNR) for each couple (Fu, Fv) of selected key frames with size (N * M), we compute the PSNR between them and the mean value is considered for each video.

$$PSNR(F_u, F_v) = 10.\log\left(\frac{N.M.255^2}{\sum_{x=1}^{N}\sum_{y=1}^{M}(F_u(x,y) - F_v(x,y))^2}\right) \tag{5}$$

4.2 Key Frame Extraction Results

In Fig. 3 we show a comparison between our proposed approach (PA) and two state of the art methods in terms of compression rate. As the CR value is high as we have different key frames. As we can see that our proposed approach (PA) reduced considerably the number of extracted key frames.

In Fig. 4, we show a comparison between our proposed approach (PA) and two state of the art methods in terms of PSNR. As the PSNR is low as we have different key frames. So, from Fig. 4 we can see that our proposed approach gives the lowest values

Table 2. Key frame extraction by the proposed method from somes standard videos "news. mpg"

Movie	Number of key frames
Filinstone.mpg	13
Foreman.avi	4
Mov1.mpg	3
HallMonitor.mpg	4
MrBean.avi	7
Coast-guard.mpg	2

Fig. 1. Key frame extraction by the proposed method from the standard video "foreman.mpg"

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

for PSNR. So, we can conclude, that it gives lowest redundancy in key frames according to CR and PSNR 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

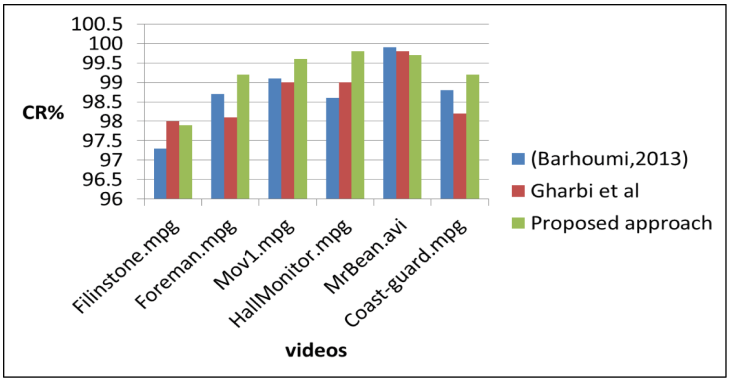

Fig. 3. Comparison of the quality of the extracted key frames in term of compression rate (CR)

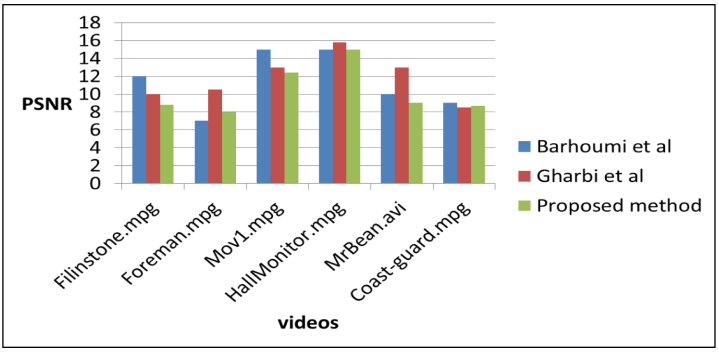

Fig. 4. Comparison of the quality of the produced results in term of PSNR values

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 shortest path algorithm (Figs. 1 and 2, Table 2).

5 Conclusions

In this paper, we have proposed a simple and effective technique for keyframe extraction based on SURF local features and using the repeatability matrix method. Firstly, candidate frames are selected adaptively using a leap extraction method. Each candidate frame is described by SURF local features vectors. Secondly, we will build the repeatability table and model it by an oriented graph. The selection of keyframes was inspired from the shortest path algorithm A*. The proposed approach proved to have superior effectiveness to others successful state of the art works, i.e., gives a set of image that covers all significant events in the standard 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 key-frames. The user can initiate his visual query by selecting one face and the system respond with videos which contains that face.

References

1. Baber, J., Satoh, S., Afzulpurkar, N. and Keatmanee, C.: Bag of visual words model for videos segmentation into scenes. In: Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, New York, NY, USA, pp. 191–194 (2013)
2. Blanken, H.M., Vries, A.P., Blok, H.E., Feng, L.: Multimedia Retrieval. Springer, Heidelberg (2010)
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
4. Ajmal, M., Ashraf, M.H., Shakir, M., Abbas, Y., Shah, F.A.: Video summarization: techniques and classification. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (eds.) ICCVG 2012. LNCS, vol. 7594, pp. 1–13. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33564-8_1
5. Uchihachi, S., Foote, J., Wilcox, L.: Automatic Video Summarization Using a Meaure of Shot Importance and a Frame Packing Method. United States Patent 6, 535,639, March 18 (2003)
6. Evangelopoulos, G., Rapantzikos, K., Potamianos, A., Maragos, P., Zlatintsi, A., Avrithis, Y.: Movie summarization based on audio-visual valiency detection. In: IEEE International Conference on Image Processing (ICIP), San Diego, CA (2008)
7. Bulut, E., Capin, T.: Key frame extraction from motion capture data by curve saliency. In: Proceedings of 20th Annual Conference on Computer Animation and Social Agents, Belgium (2007)
8. Peyrard, N., Bouthemy, P.: Motion-based selection of relevant video segments for video summarization. Multimedia Tools Appl. **26**(3), 259–276 (2005)
9. Li, C., Wu, Y.T., Yu, S.S., Chen, T.: Motion-focusing key frame extraction and video summarization for lane surveillance system. In: 16th IEEE International Conference on Image Processing (ICIP), pp. 7–10 (2009)
10. Chheng, T.: Video Summarization Using Clustering. Department of Computer Science University of California, Irvine (2007)
11. Damnjanovic, U., Fernandez, V., Izquierdo, E.: Event detection and clustering for surveillance video summarization. In: Proceedings of the Ninth International Workshop on Image Analysis for Multimedia Interactive Services. IEEE Computer Society, Washington (2008)
12. Liu, D., Chen, T., Hua, G.: A hierarchical visual model for video object summarization. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 2178–2190 (2010)
13. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2012)
14. Calic, J., Thomas, B.: Spatial analysis in key-frame extraction using video segmentation. In: Proceedings of Workshop Image Analysis of Multimedia Interactive Services Lisbon, Portugal (2004)

15. Calic, J., Izquierdo, E.: Efficient key-frame extraction and video analysis. In: Proceedings of International Conference on Information Technology: Coding and Computing, pp. 28–33 (2002)
16. Liu, X., Song, M.L., Zhang, L.M., Wang, S.L.: Joint shot boundary detection and key frame extraction. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR 2012), pp. 2565–2568 (2012)
17. Ejaz, N., Tariq, T.B., Baik, S.W.: Adaptive key frame extraction for video summarization using an aggregation mechanism. J. Vis. Commun. Image Represent. **23**, 1031–1040 (2012)
18. Xu, Q., Liu, Y., Li, X., Yang, Z., Wang, J., Sbert, M., Scopigno, R.: Browsing and exploration of video sequences: a new scheme for key frame extraction and 3D visualization using entropy based Jensen divergence. Inf. Sci. **278**, 736–756 (2014)
19. Lai, J.L., Yi, Y.: Key frame extraction based on visual attention model. J. Vis. Commun. Image Represent. **23**, 114–125 (2012)
20. Kumar, M., Loui, A.C.: Key frame extraction from consumer videos using sparse representation. In: Proceedings of the 18th IEEE International Conference on Image Processing (ICIP 2011), pp. 2437–2440, (2011)
21. Chergui, A., Bekkhoucha, A., Sabbar, W.: Video scene segmentation using the shot transition detection by local characterization of the points of interest. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 404–411 (2012)
22. Tapu, R., Zaharia, T.: A complete framework for temporal video segmentation. In: 2011 IEEE International Conference on Consumer Electronics-Berlin (ICCE-Berlin), pp. 156–160 (2011)
23. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
24. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. Int. J. Comput. Vis. **37**, 151–172 (2000)
25. Gharbi, H., Bahroun, S., Zagrouba, E.: A novel key frame extraction approach for video summarization. In: International Joint Conference on Computer Vision Theory and Applications, Rome (2016)
26. Barhoumi, W., Zagrouba, E.: On-the-fly extraction of key frames for efficient video summarization. In: AASRI Conference on Intelligent Systems and Control (2013)
27. Bo, C., Lu, Z., Dong-ru, Z.: A study of video scenes clustering based on shot key frames. Wuhan Univ. J. Nat. Sci. **10**, 966–970 (2005). Series Core Journal of Wuhan University (English)
28. Sargent, G., Perez-Daniel, K.R., Stoian, A., Benois-Pineau, J., Maabout, S.: A scalable summary generation method based on cross-modal consensus clustering and OLAP cube modeling. Multimedia Tools Appl. **75**, 1–22 (2016)