Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 4 (2013) 78 – 84

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Intelligent Systems and Control

On-the-fly extraction of key frames for efficient video summarization

Walid Barhoumi and Ezzeddine Zagrouba

Research Team "Systèmes Intelligents en Imagerie et Vision Artificielle" (SIIVA) – RIADI Laboratory, ISI, 2 Rue Abou Rayhane Bayrouni, 2080 Ariana, Tunisia

Abstract

We propose in this paper an object-based method for on-the-fly extraction of key frames summarizing the salient visual content of videos. This method is based on spatial segmentation of each frame in order to detect the important events. Thus, key frame detection is facing a much more semantic criterion so that each key frame presents an important event such as the appearance and the disappearance of relevant objects. Realized experiments on challenging videos demonstrate the efficiency of the proposed method, which is able to catch the semantic content of a video shot, while preventing redundancy of the extracted key frames and maintaining minimum requirements in terms of memory space.

*Keywords:*region segmentation; key frame extraction; video summary; many-to-many assignment; on-the-fly.

1. Introduction

Nowadays users are faced with an ever-growing amount of videos and efficient tools are thus more and more needed for videos archiving, indexing and retrieval. Video summarization is an important research topic which aims to create automatically a compact and representative summary of video content in terms of still images. Most of existing video summarization methods are content-based, since keyword-based ones pose severe problems of subjectivity and feasibility. Moreover, in order to overcome the "semantic gap", between the description of a video in terms of low-level features and its semantic content, recent methods combine low-level criteria with the semantic notion of objects. There are two classes of video summarization methods based on objects. For the first group, the concept of object is used to extract the most representative frames (key frames) which represent the salient content and the second group includes methods providing background-foreground segmentation [3]. We focus our attention here on video summarization with a minimum amount of data by extracting relevant key-frames. However, key frame extraction has strong limitations in the case of long

doi:10.1016/j.aasri.2013.10.013

videos. In fact, much of the visual content of the frames is redundant and/or irrelevant, and it is necessary to retain only the information strictly needed for functional browsing and querying. Thus, it is preferable to divide the input video into shots, and each shot will be then represented by some key frames. Earlier key frame extraction methods are based on temporal sampling while uniformly choosing key frames at predefined intervals [4]. These approaches are not content-based and do not consider dynamics of the visual content and selected frames are often unstable. However, content-based methods can be grouped into three classes. Methods belonging to the first class are based on frames clustering. The idea is to group the frames into homogenous clusters and key frames are then selected from each cluster or only from the largest ones. In fact, frames composing a cluster should share the visual content and the closest frame to the centroid represents usually a recurring visual content. To evaluate the similarity between two frames, color histograms and local color averages can be used [6]. Other methods classify frames while considering regions of interest (key-objects) of each frame [11]. The main drawback of these methods is that, depending on the number of clusters, key frames can be either redundant or fail to represent efficiently the content of the whole shot [9]. Besides, they are computationally intensive and do not consider the temporal information [5]. The second class includes methods based on statistical analysis. Earlier methods assume that a frame should be selected as key frame if it is different from the previous frame. Park et al., 2005 proposed to estimate the coverage rate of each frame, such that frames maximizing this rate are considered as candidate key frames [10]. Then, distortion rate permits to select final key frames among candidates. Nevertheless, it is ambiguous to select which frame is the best key frame when several candidates have high coverage rates. Other methods extract key frames while comparing non-adjacent frames, using inter-frames entropy, histograms similarity or wavelets [8]. In order to integrate a more semantic concept, recent methods are based on statistical models applied on key-objects. Sun and Ping, 2004 select only frames with the maximum ratio of objects to background [12]. However, methods forming this class do not prevent redundancy of key frames and do not handle correctly the loop closure cases. Moreover, most of these methods are threshold-depending and impose that the number of key frames is set a priori. The third class includes methods based on camera motion analysis. They assume that key frames are focused by the camera. For this, a curve illustrating the evolution of motion amplitude across the shot is often generated and local extrema define key frames. To estimate camera motion, frames difference, optical flow, and blocks matching can be used [7]. Methods belonging to this class are widely used when dealing with compressed videos, since they express the dynamics of a shot through motion analysis. However, in addition to the extensive computation time of these methods, the underlying assumption that key frames correspond to the local extrema of the camera motion is not necessarily correct [10].

In this paper we combine object segmentation with low-level features in order to propose a higher level of description in terms of semantic primitives. The visual content of a shot is summarized on-the-fly into key frames, such that each one represents a new event. The first frame of the shot is automatically selected as key frame. Then, each received frame is segmented into salient objects, and a position-based criterion is combined with a shape-based one to reject irrelevant objects. Next, many-to-many correspondence between objects of the current frame and those extracted in previous key frames allows to decide if the current frame corresponds to a new event. The main contribution of the proposed method is the summarization on-the-fly of a shot while extracting key frames illustrating relevant events. Many tests on standard videos showed that the proposed method is able to preserve the overall content of a shot with minimum data, even when the camera returns to parts of the scene already visited before. Next section describes the proposed method and experimental results are presented in section 3 to prove objectively the effectiveness of the method using standard metrics.

2. Proposed method

The proposed method for key frames extraction is mainly based on shot boundary detection and object-based event detection. In fact, after partitioning the input video into shots [2], key frames are selected on-the-fly from each shot while looking for important events corresponding to the appearance and the disappearance of significant objects. For that purpose, the first frame F_1 in each input shot is automatically considered as key frame KF_1 and is also segmented into salient objects using a fuzzy coarse region segmentation technique [1].

This fuzzy segmentation technique consists of applying the watershed algorithm followed by a region-growing process which merges seeded watershed regions according to their histograms similarities, in order to overcome the usual over-segmentation effects of watersheds. Besides, a post-treatment keeps only relevant objects, using two criteria. The first (*resp.* second) criterion is based on the assumption that if an object is important then the camera will focus on it (*resp.* it has a compact shape). Thus, only objects located around the middle of the frame with low compactness are considered as relevant (Fig. 1). The position-based criterion rejects inaccurate regions on the borders, and the shape-based one excludes the dispersed regions, mainly thin and elongated ones, which are due to the leak from some foreground objects to the background during the segmentation (Fig. 1). Then, given the set ζ_1 of relevant objects belonging to KF_1, the same object extraction process is applied on each received frame F_t of the shot, in order to detect the relative set ζ_t of relevant objects and to decide then if F_t can be considered as key frame or not. Indeed, if the frame F_t corresponds to the appearance or to the disappearance of one, or more, significant objects, comparatively to ζ_1, then F_t is considered as a new key frame and should be added to the set χ of the resulting key frames. Formally, given the sets ζ_1 and ζ_t of the accurate objects belonging to KF_1 and KF_t respectively, the frame F_t represents a new event only if it exists at least one object in ζ_1 (*resp.* in ζ_t) which cannot be matched with any object in ζ_t (*resp.* in ζ_1). To match an object O_t^i ($\in \zeta_t$) with another one O_1^j ($\in \zeta_1$), we used a model of appearance based on the local color histograms, which is insensitive to the background noise and to the camera view-point change. Thus, two salient objects O_t^i and O_1^j belonging to two different frames are matched only if their similarity degree *Sim* (1) is close to 1. The similarity of two objects is based on the resemblance of the visual content of the four blocks composing each object. As the color histogram does not integrate spatial information, we used it as a local descriptor within each block, such that these blocks are defined according to the principal factorial axes of the object. Thus, visual similarity of each couple of blocks ($\in O_t^i \times O_1^j$) is evaluated in terms of the intersection of the correspondent 16×8 HS histogram in the HSV color space, which is often regarded as the closest color space to the human visual perception.

$$Sim(O_t^i, O_1^j) = \sum_{b=1}^{4} \sum_{h=1}^{16} \sum_{s=1}^{8} \frac{\min(hist_t^i(h,s,b), hist_1^j(h,s,b))}{\max(hist_t^i(h,s,b), hist_1^j(h,s,b))}. \quad (1)$$

Fig. 1. Object extraction. The first image is the original frame and the following ones are the extracted objects: only the 2nd region is considered as relevant, since the 3rd and 4th ones (*resp.* 1st) are excluded by the position-based criterion (*resp.* the shape-based criterion).

Once the dissimilarity between each couple (O_t^i, O_1^j) of relevant objects ($\in \zeta_t \times \zeta_1$) is defined, objects' correspondence between the correspondent frames is formulated as a many-to-many linear assignment problem between ζ_t et ζ_1. To resolve this problem, we applied the shortest augmenting path algorithm while looking to find the association between the objects that maximize the amount of the corresponding similarities. This many-to-many assignment permits to handle correctly complex interactions and occlusions between objects, without considering these situations as new events. The many-to-many correspondence between objects allows the comparison of the objects at different levels of granularity what overcomes the over- and under-segmentation effects. If more than one key frame is already extracted ($Card(\chi)>1$), a received frame is assumed to represent a new key frame only if it cannot be associated with all key frames already extracted up to that instant. Indeed, given the set $\chi=\{KF_1,...,KF_j\}$ of the already extracted key frames, the frame F_t is assumed to represent a key frame ($\chi \leftarrow \chi \cup F_t$) only if it represents a new event comparatively to KF_j, to KF_{j-1}, . . .and to

KF_1, while giving priority to recent key frames (initially, $ref=j=Card(\chi)$). Thus, F_t is considered as key frame only if some objects are totally added to, or removed from, the scene in this frame (2). In addition to the implicit integration of the temporal behavior with the visual appearance of salient objects, the comparison of each frame with the already selected key frames avoids the consideration of the temporally appearance or disappearance of objects as new events. In particular, the given priority to recent key frames allows the detection of the repetitive events without testing a huge set of frames. Besides, the many-to-many assignment avoids the consideration of occlusion effects as new events, what minimizes the redundancy of the final key frames. Once the decision is performed for F_t, the same object-based event detection procedure is applied on-the-fly for the next received frame F_{t+1} and so on until the end of the shot. We note that from the second shot, each frame F_t, including the first one of the shot, must be compared not only to the key frames already extracted from this shot but also to those selected within all precedent shots.

$$F_t \text{ represents a new event relatively to } F_{ref} \Leftrightarrow \underbrace{\left[\exists O_t^i \in \xi_t \,/\, \max_{O_{ref}^j \in \xi_{ref}} Sim(O_t^i, O_{ref}^j) \le \theta\right]}_{\text{appearance of objects}} \text{OR} \underbrace{\left[\exists O_{ref}^j \in \xi_{ref} \,/\, \max_{O_t^i \in \xi_t} Sim(O_t^i, O_{ref}^j) \le \theta\right]}_{\text{disappearance of objects}}. \quad (2)$$

3. Experimental results

The proposed key frame extraction method was applied on several videos (news, cartoons, games,…) illustrating different challenges (camera motion, background-foreground similar appearance, dynamic background,…). Results prove that the method is able to extract efficiently few key frames resuming the salient semantic content of a video (Fig. 2). Notice that in the case of a reduced number of key-objects within the input video, our method extracts only relevant and non redundant key frames. For example, only 3 relevant key frames were extracted from "mov1.mpg", which is composed of 377 frames partitioned into 9 shots. Indeed, in many cases, even the first frame of a shot is not necessary a key frame since it represents an old event already selected among the precedent shots. Thus, the precision of the proposed method is equals to 100% for these videos, and this value is clearly higher in comparison to those recorded with other methods [14]. However, for complex videos, the method performance decreases slightly because of the complexity of the appearance and the objects interaction (Fig. 2). For example, in "flinstone.mpg", the background color is very similar to this of the moving objects, what can explain the redundancy of some key frames. In particular, the moving objects in this video interact extensively with long partial and total occlusion effects. Thus, 14 key frames are detected in this video, which is preliminary subdivided into 13 shots. This is mainly due to the significant variation of the visual content of the frames composing some shots. For the video "bmw.avi" which is very challenging since the visual content changes rapidly in temporal scale, the proposed method extracts only 9 key frames in 9 shots. The redundancy of some successive key frames for this video is mainly due to the extensive camera motion, particularly zooming effects, which affected even the shots boundary detection. We note that although the camera reverses direction and loops back revisiting some parts of the scene in "mov1.mpg" and "bmw.avi", only few relevant and non redundant key frames are selected.

We also evaluated objectively the quality of the produced results while using various standard quality measures [15]. We compared the resulting quality measures (Fig. 3) by the proposed method (PM) with those recorded with five competing state-of-the-art methods [10, 14, 8, 13, 2] and this for eleven standard test videos. These methods were selecting so that they represent a sampling of the existing classes of key frame extraction methods (c.f. Section 1). According to the compression ratio (CR), it is clear that the proposed method minimizes considerably the redundancy of the extracted key frames, what guarantees encouraging compression ratios while maintaining minimum requirements in terms of memory space. For example, while only 9 key frames are extracted from 9 shots composing "bmw.avi", the other methods select much more key frames with significant redundancy in their semantic content. We note that the average compression ratio provided by our method is around 98.6% with the lowest standard deviation (0.0091), comparatively to the compared methods (Fig. 3.a). On the other side, for an objective assessment of the quality of the extracted key frames, we used the

signal to noise ratio (PSNR). In fact, for each couple (F_u, F_v) of selected key frames (of size $N \times M$), we measure the PSNR between them (3) and the mean value is recorded for each studied video (Fig. 3.b). 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. 3.b). These values confirm that our method extracts the most significant and relevant key frames while minimizing redundancy. Besides, we evaluated the semantic pertinence of the extracted key frames while carrying out a subjective test where 8 junior researchers on computer vision assess independently the satisfaction level of each key frame extraction method for the eleven test videos. The average precision value of the proposed method (≈73%) is much higher than those of the compared ones (≈54%), what confirms the accuracy of the selected key frames by our method.

$$PSNR(F_u, F_u) = 10.\log\left(\frac{N.M.255^2}{\sum_{x=1}^{N}\sum_{y=1}^{M}\left(F_u(x,y) - F_v(x,y)\right)^2}\right), \quad (3)$$

Fig. 2. Key frame extraction by the proposed method from standard videos.

4. Conclusion

We presented an efficient object-based method for key frame extraction while maintaining convenient memory requirements, even under loop-closure situations. This method is mainly based on the detection of significant events while analyzing the spatio-temporal behavior and visual appearance of salient objects. The detection on-the-fly of the key frames allows to integrate implicitly the temporal content within the input shot without having to process the whole shot. This allows a properly capturing of the underlying dynamics of the input frames. To the best of our knowledge, not much attention was paid for the realization of this task on-the-fly. Moreover, the suggested method avoids the complexity of existing methods based on clustering or optimization strategies. It captures efficiently the underlying dynamics of frames without requiring a priori knowledge of the number of frames representing each shot. The preliminary recorded results and an objective comparative study with many existing key frame extraction methods have shown the efficiency of our unsupervised content-based method, in terms of redundancy, compression rate and recall/precision imetrics. As perspectives, we propose to use a visual dictionary, which can be formed based on low-level attributes of the relevant objects of interest in all available frames, in order to provide a model vector that describes each frame based on the types of objects it contains.

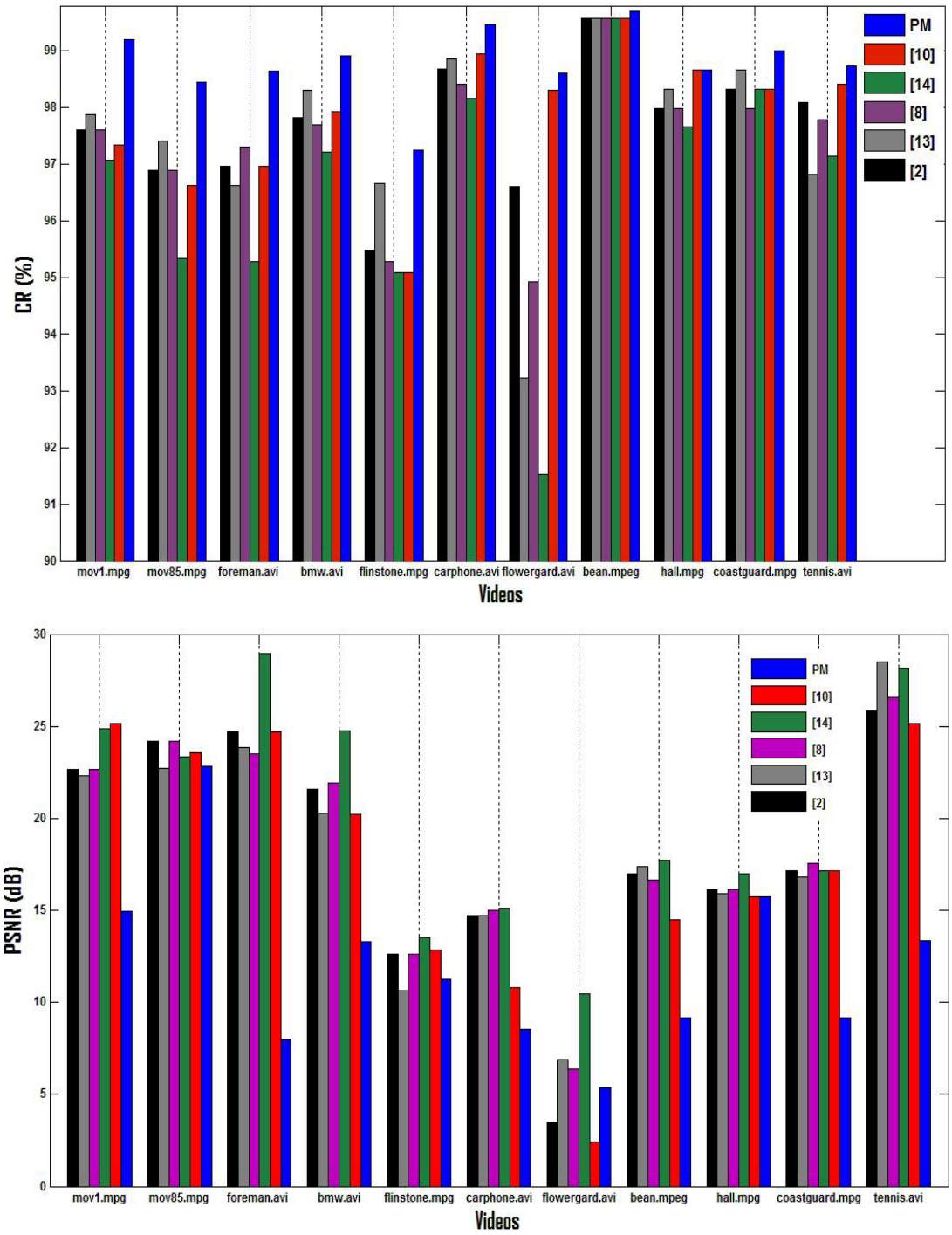

Fig. 3. Objective comparison of the quality of the produced results, for eleven test videos, by the proposed method (PM) with those recorded with five state-of-the-art methods while using compression ratio (CR) and PSNR.

References

[1] Barhoumi W, Gallas A, Zagrouba E. Effective region-based relevance feedback for interactive content-based image retrieval. Studies in Computational Intelligence 2009; 226:177-187.

[2] Bo C, Lu Z, and Dong-Ru Z. A study of video scenes clustering based on shot key frames. Natural Sciences 2005; 10:966-970.

[3] Amri S, Barhoumi W, Zagrouba E. A robuste framework for joint background/foreground segmentation in complex video scenes filmed wit freely moving camera. Multimedia Tools and Applications 2010; 46:175-205.

[4] Lee HC, Kim, SD. Iterative keyframe selection in the rate-constraint environment. Sig Proc Imag Com 2003; 18:1-15.

[5] Li Z, Schuster G, Katsaggelos AK, Gandhi B. Optimal video summarization with a bit budget constraint. ICIP 2004; 613-616.

[6] Liu D, Shyu ML, Chen CC, Chen SC. Integration of global and local information in videos for key frame extraction. Int Conf on Information Reuse and Integration 2010; 171-176.

[7] Liu G, Zhao J. Key frame extraction from MPEG video stream. Int Symp on Information Processing 2010; 423-427.

[8] Mentzelopoulos M, Psarrou A. Key-frame extraction algorithm using entropy difference. ACM Int Work on Multimedia Information Retrieval 2004; 39-45.

[9] Mukherje S, Mukherje P. A design-of-experiment based statictical technique for detection of key-frames. Multimedia Tools and Applications 2013; 62:1-31.

[10] Park KT, Lee JY, Rim KW, Moon YS. Key frame extraction based on shot coverage and distortion. LNCS 2005; 3768:291-300.

[11] Spyrou E, Tolia G, Mylonas P, Avrithis Y. Concept detection and keyframe extraction using a visual thesaurus. Multimedia Tools and Applications 2009: 41:337-373.

[12] Sun Z, Ping F. Combination of color and object outline based method in video segmentation. SPIE 2004; 5307:61-69.

[13] Wolf W. Key frame selection by motion analysis. Int Conf on Acoustic, Speech and Signal Processing 1996; 1228-1231.

[14] Zhuang Y, Rui Y, Huang TS, Mehrotra S. Adaptive key frame extraction using unsupervised clustering. ICIP 1998; 866-870.

[15] Ejaz N, Bin Tariq T, Baik SW. Adaptive key frame extraction for video summarization using an aggregation mechanism . Visual Communication and Image Representation 2012; 23: 1031-1040.