NON-SEQUENTIAL VIDEO CONTENT REPRESENTATION USING TEMPORAL VARIATION OF FEATURE VECTORS

Anastasios D. Doulamis, Nikolaos Doulamis, Georgios Akrivas and Stefanos Kollias
Electrical and Computer Engineering Department National Technical University of Athens
Zografou 15773, Athens, Greece E-mail: adoulam@image.ntua.gr

ABSTRACT

In this paper, an efficient non-sequential representation of the video content is presented based on the temporal variation of the extracted feature trajectory. The proposed scheme is very fast and accurate and can be applied to any generic video stream. Many benefits in the filed of image communication, storage requirements and efficient performance of indexing and retrieval algorithms can gain from such a non-linear representation.

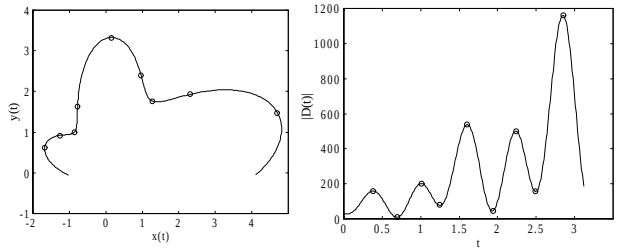
INTRODUCTION

The traditional representation of video as a sequence of numerous consecutive frames, each of which corresponds to a constant time interval (40ms for the PAL system), stems from the analog tape storage process and results in a linear (sequential) access of video content. While this approach is adequate for viewing a video in a movie mode, it has a number of limitations for the new emerging multimedia applications, such as video browsing, content-based indexing and retrieval. Furthermore, such video representation is not adequate for efficient organization of large video archives since storage requirements of digitized video information, even in compressed domain, are very large. To overcome the aforementioned difficulties a non-sequential (non-linear) video content representation should be applied using a content-based sampling algorithm to extract a small but meaningful information of visual content.

In this paper, a new fast and efficient algorithm for non-sequential video content representation is proposed applicable to any generic video sequence. Instead, most of the previous works use complicated techniques, such as construction of a compact image map or image mosaics [1], [2], or extract a simple key-frame at constant time instances [3]. The former methods, however, are also restricted to specific applications and cannot provide good results in case of real-world scenes, where background/foreground changes or complicated camera effects may appear. On the other hand, the latter cannot describe the visual content with high efficiency since it is possible important shots of small duration to have no representatives while shots of longer duration to be represented by multiple frames with similar content.

THE PROPOSED SCHEME

The first stage of the proposed algorithm is to transform the traditional pixel-based representation of an image to a feature-based one. This is accomplished by extracting several descriptors about the visual content. In our case, two different groups of descriptors are used. The first refers to the global visual image characteristics, like, for example, the global color or motion histogram. The second exploits object-based properties. In our case, video objects are extracted by applying segmentation algorithm both in color and motion domain. To reduce the required computational cost, a multiresolution implementation of the *Recursive Shortest Spanning Tree* (RSST) algorithm, called *M-RSST*, has been also implemented. By comparing, the computational cost required for the adopted M-RSST with the cost of the conventional RSST using a C implementation on a Sun SparcStation-20 system it can be seen that the improvement ratio is up to 400 times for images of 720x576 pixel size. Then, all the aforementioned features are gathered together, using a fuzzy feature vector formulation as in [4].



(a) (b)

Fig.1: (a) A continuous curve r(t)=(x(t),y(t)), and (b) the magnitude of the second derivative D(t) versus t.

Key-frames are extracted in the following, based on the temporal variation of feature vectors. In particular, the feature vectors for all frames within a shot form a trajectory and thus selection of the most representative frames within a shot is equivalent to selection of appropriate curve points, able to characterize the corresponding trajectory. This can be achieved by extracting the time instances, i.e., frame numbers, which reside in extreme locations of this trajectory. The magnitude of the second derivative of the feature vector with respect to time is used as a curvature measure in this case. In particular, the most representative frames are selected based on the maximum and minimum of this magnitude. This is due to the fact that local maxima correspond to time instances of peak variation of object velocity, while local minima to almost constant velocity. For example, suppose that we have a 2-dimensional feature vector whose trajectory is illustrated in Fig. 1 as a continuous curve $\mathbf{r}(t)=(x(t),y(t))$. Then the local maxima and minima of the magnitude of the second derivative $D(t)$, shown as small circles in Fig. 1(a,b), do provide sufficient

information about the curve shape, since it can be reproduced using some kind of interpolation.

The proposed scheme has been applied to real-life video sequences. Fig. 2 illustrates an example, coming from a test drive sequence and consisting of N_s=223 frames. One every 10 frames is depicted, resulting in 23 frame thumbnails. The results on this shot are presented in Fig. 3. In particular, Fig. 3(a) shows the magnitude of the second windowed derivative, $|D(k)|$, versus the frame number, k. To eliminate possible noise on the curve, the feature trajectory has been first smoothed, by applying a low pass filter. The 7 selected frames are depicted in Fig. 3(b). It can be seen that these frames provide sufficient visualization of the total 223 frames of the shot.

CONCLUSIONS

Key-Frame extraction based on temporal feature vector variation is an extremely fast and very straightforward algorithm since, in discrete time, the second derivative is implemented as a difference equation. In addition, the number of key frames for a shot is not required to be a priori known. Instead, it is estimated by the feature vector trajectory.

Such non-sequential video representation presents many advantages both in the field of image processing and communication. A) It reduces storage requirements. For example, for a 30-min video (consisting of 200 shots), instead of storing about 45,000 frames, it can be stored only 1,000 if 5 key-frames are selected per shot and. B) It permits easy video transmission over IP protocols since the video thumbnails (key-frames) can be first transmitted in contrast to the entire sequence. C) It provides efficient indexing and retrieval on video database, since it reduces all the redundant information.

REFERENCES

[1] M. Irani and P.Anandan, "Video Indexing Based on Mosaic Representation," *Proceedings of the IEEE*, Vol. 86, No. 5., pp. 805-921, May 1998.

[2] N. Vasconcelos and A. Lippman, "A Spatiotemporal Motion Model for Video Summarization," *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 361- 366, Santa Barbara, USA, June 1998.

[3] M. Mills, J. Cohen and Y. Y. Wong, "A Magnifier Tool for Video Data," *Proc. ACM Computer Human Interface (CHI)*, May 1992, pp. 93-98.

[4] A. D. Doulamis, N. D. Doulamis and S. D. Kollias, " A Fuzzy Video Content Representation for Video Summarization and Content-Based Retrieval," *Signal Processing*, Elsevier Press, March 2000.

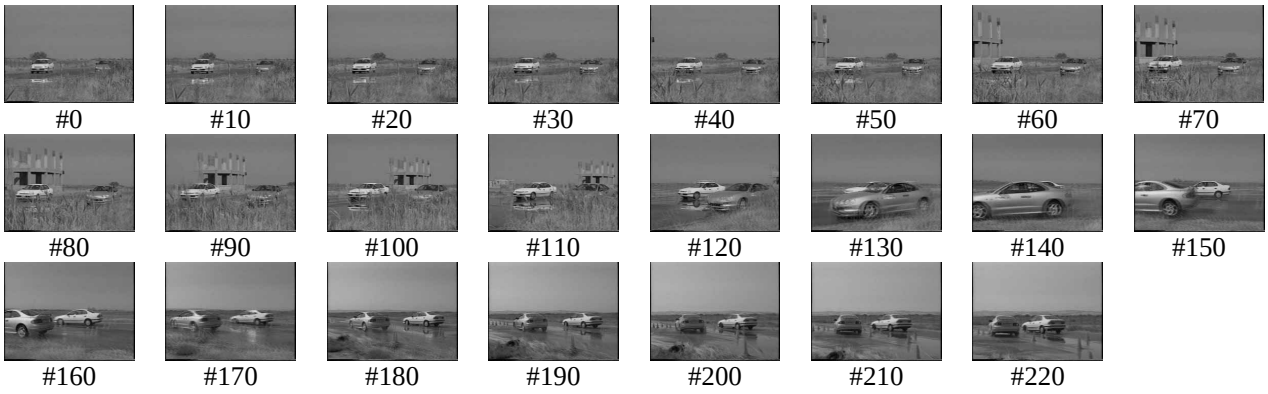


Fig.2: Test drive sequence, frames #0 to #220.

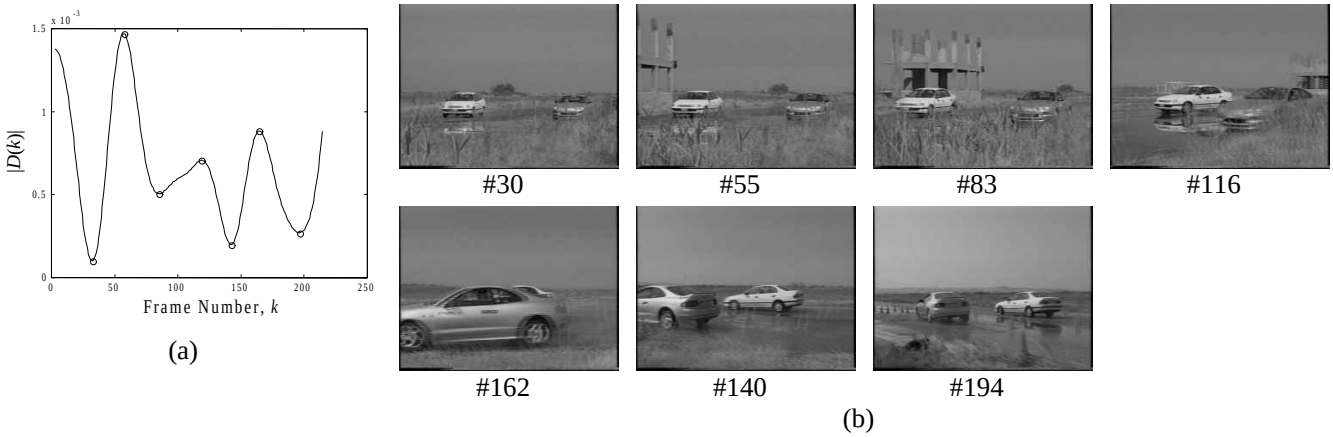


Fig.3: Temporal variation approach on test drive sequence: (a) magnitude of second windowed derivative, |D(k)|, versus the frame number, k, and (b) selected frames.