Shot Boundary Detection Techniques for Video Sequences

H. Koumaras
Institute of Informatics and Telecommunications, Greece

G. Xilouris
Institute of Informatics and Telecommunications, Greece

E. Pallis
Technological Educational Institute of Crete, Greece

G. Gardikis
University of the Aegean, Greece

A. Kourtis
Institute of Informatics and Telecommunications, Greece

INTRODUCTION

The advances in digital video encoding and compression techniques that achieve high compression ratios by exploiting both spatial and temporal redundancy in video sequences have made possible the storage, transmission, and provision of very high-volume video data over communication networks.

Today, a typical end user of a multimedia system is usually overwhelmed with video collections, facing the problem of organizing them in a browsing-friendly way. Thus, in order to allow an efficient exploitation and browsing of these video-anthologies, it is necessary to design techniques and methods for content-based search and access. Therefore, the issue of analyzing and categorizing the video content by retrieving highly representative optical information has been raised in the research community.

Thus, the current trend has led to the development of sophisticated technologies for representing, indexing, and retrieving multimedia data. A common first step towards this is the segmentation of a video sequence into elementary shots, each comprising a sequence of consecutive frames that record a video event or scene continuously in the spatial and temporal domain. Moreover, these elementary shots appear as they have been captured by a single camera action. Two adjacent elementary streams are divided by a *shot boundary* or *shot transition,* also known as scene cut, when the change of video content occurs over a single frame, or *gradual shot boundary,* when the changes occur gradually over a short sequence of frames (e.g., dissolve, fade in/out, etc.) (Lu & Tan, 2005).

In general, gradual transitions are more demanding in detection than abrupt scene cuts, because they must be distinguished from regular camera operations that cause similar temporal variances and usually trigger false detections. Especially for video content with high spatial and temporal activity level, the detection of gradual scene changes becomes even more challenging. (Hampapur, Jain, & Weymouth, 1995).

Hence, the goal of this temporal video segmentation is to divide the video stream into a set of meaningful and manageable segments that are used as basic elements for indexing. Further analysis may be performed, such as representation of the video content and event identification.

In future multimedia systems, the offered video services will be provided in the form of MPEG-21 digital items, which integrate a typical encoded media clip along with its XML-based metadata descriptors, enabling in this way advanced search and retrieve abilities. Also future multimedia implementations will adapt MPEG-21 schema, which means that upcoming media recorders must be able to automatically create video content indexing.

This chapter will outline the various existing methods of boundary shot and scene change detection.

BACKGROUND

A primitive typical approach to indexing video data was the manual creation of textual annotations along with time headers in the metadata of a media file. However, such a human-based method is time consuming and practically not applicable. Moreover, such methods suffer from the subjectivity of the human operator during the textual description.

Therefore, it is necessary to develop an integrated framework for automatic extraction of the most characteristic frames of a video sequence, which will finally enable the efficient indexing and description of a video sequence. More specifically, by developing methods that enable the automatic build of a scene-access menu for a video clip, the viewer may use this index for quick access at a specific scene or for performing scene searches.

Several approaches have been proposed in the literature for automatic video indexing, which can be basically categorized as methods for temporal segmentation in an uncompressed or compressed video domain (Koprinska & Carrato, 2001; Lienhart, 1999; Dailianas, Allen, & England, 1995).

Thus, the various temporal video segmentation methods for each class (i.e., uncompressed/compressed) will be discussed in the following sections.

SHOT BOUNDARY DETECTION IN UNCOMPRESSED DOMAIN

Video segmentation in an uncompressed domain includes all the boundary shot detection methods that perform using metrics and mathematical models on the uncompressed/spatial video signal. Most existing methods detect shot boundaries of video based on some change of the video content on the visual domain between consecutive frame pairs. If the measured change is above a predetermined threshold, then a shot boundary is assumed and reported.

Based on the metrics nature that is used to detect the differences between successive frames, the algorithms can be generally classified into the following classes: pixel-based, block-based, and histogram-based (Zhang, Low, Gong, & Smoliar, 1994, 1995).

Pixel-Based Methods

Pixel-based methods evaluate the differences in luminance or color domain between pixel values of successive frames (Kikukawa & Kawafuchi, 1992). Hence, a per pixel comparison is performed between frame pairs. Depending on the measured difference from the pixel-based comparison, a scene cut is detected and reported if the calculated difference is above a pre-defined threshold value. Otherwise no scene change is reported. The sensitivity and the efficiency of the pixel-based methods are strongly related to the selection of the reference threshold.

Block-Based Methods

In contrast to the aforementioned pixel-based methods, where the whole frame of a video movie is taken under consideration for the scene change detection and the corresponding measured difference in the pixel values, either in color or luminance domain, in block-based methods each frame is divided into blocks that in turn are compared to their corresponding blocks in the successive frame (Kasturi & Jain, 1991; Shahraray, 1995). More specifically, in contrast to the aforementioned pixel-based techniques, where the critical unit is the number of pixels whose difference is above a threshold value, these methods report a scene change, when the number of changed blocks is greater than a predefined threshold.

Histograms Comparisons

The aforementioned categories exploit pixel comparison in order to derive a decision. On the contrary, histogram-based methods exploit the fact that a set of frames that belong in the same scene retain, in general unchanged, their luminance- or color-level histograms. A luminance- or color-level histogram of a frame depicts the density of the number of pixels that have specific luminance or color value.

As has been described, the majority of the aforementioned methods are implemented based on metrics of the uncompressed video domain, utilizing a common framework: a similarity measurement between successive frames.

SHOT BOUNDARY DETECTION IN COMPRESSED DOMAIN

Multimedia applications that distribute audiovisual content over communication networks (such as video-on-demand (VOD) and real-time entertainment streaming services) are based on digital encoding techniques (e.g., MPEG-1/2/4 and H.261/2/3 standards) that achieve high compression ratios by exploiting the spatial and temporal redundancy in video sequences. Most of the standards are based on motion estimation and compensation, using the block-based discrete cosine transformation (DCT). The use of transformation facilitates the exploitation in the compression technique of the various psychovisual redundancies by transforming the picture to a domain where different frequency ranges with dissimilar sensitivities at the human visual system (HVS) can be accessed independently.

The DCT operates on a X block of N X N image samples or residual values after prediction and creates Y, an N X N block of coefficients. The action of the DCT can be described in terms of a transform matrix A. The forward DCT is given by:

$$Y=AXA^T$$

where X is a matrix of samples, Y is a matrix of coefficients, and A is an N X N transform matrix. The elements of A are:

$$A_{ij} = C_i \cos\frac{(2j+1)i\pi}{2N},$$

where

$$C_i = \begin{matrix} \sqrt{1/N}, i = 0 \\ \sqrt{2/N}, i > 0 \end{matrix} \qquad (1)$$

Therefore the DCT can be written as:

$$Y_{xy} = C_x C_y \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} X_{ij} \cos\frac{(2j+1)y\pi}{2N} \cos\frac{(2i+1)x\pi}{2N} \qquad (2)$$

Afterwards in the encoding chain, quantization of the aforementioned DCT coefficients is performed, which is the main reason for the quality degradation and the appearance of artifacts, like the ‘blockiness’ effect.

Several methods for shot boundary detection in the compressed domain have been developed. According to Koprinska and Carrato (2001), they can be classified into the following categories, depending on the used metric:

- **DCT Coefficients:** The temporal video segmentation methods based on DCT coefficients apply a comparison technique to the DCT coefficients of the corresponding successive video frames. The difference metric is somewhat similar to the aforementioned pixel-based metric, where a scene change is detected and reported when the measured difference exceeds a specific threshold value (Zhang et al., 1994, 1995).
  It must be noted that these methods can be applied only on intra-coded frames of a DCT-based coded signal, because only they are fully encoded with DCT coefficients. Thus, the processing requirements may be low, but the temporal accuracy of the detected frame drops dramatically, and it is highly dependent on the intra-frame periodicity.
- **DC Terms:** The DC term is a scaled version of the average value for each block and thus the DC terms are directly related to the pixel domain. So, in a similar way to the uncompressed domain methods, the DC terms-based metrics measure the DC terms differences between successive frames. Again a frame is reported as shot boundary, if the aforementioned measurement is higher than a pre-defined threshold (Yeo & Liu, 1995).
- **DC Terms, Macroblock (MB) Coding Mode:** This is a hybrid method in which, except from the aforementioned DC terms, the type of the macroblock (MB) coding is taken under consideration as well. When a scene change takes place, then some macroblocks of an inter-coded frame may be intra-coded due to limited reference options, demonstrating where scene change occurs (Meng, Juan, & Chang, 1995).
- **MB Coding Mode and Motion Vectors (MVs):** Similarly to the previously described method, a hybrid model is exploited, where it takes under consideration both the MB coding mode and the MV information of the encoded video sequence.
- **MB Coding Mode and Bit Rate Information:** Finally this method uses both bit rate information and motion-predicted MB types in order to derive accurate estimation of the scene changes. The forced intra-coding of some MBs over a scene change increases the deduced bit rate due to the inefficiency of the intra-coding method.

Similarly to the shot change detection methods of the uncompressed domain, this section has shown that also the methods of the compressed domain exploit analogous frameworks at their implementation, which is based on the comparison of the calculated metric between successive frame pairs.

FUTURE TRENDS

All the aforementioned methods use a threshold parameter in order to distinguish shot boundaries and changes. Thus, a common problem in shot boundary detection lies in the selection of an appropriate threshold for identifying whether a change is sufficiently large to signify a shot boundary or not (Lu & Tan, 2005). If a global threshold is used for the detection of shot boundaries over the whole video, then successful detection rate may vary up to 20% even for the same video content (O’Toole, Smeaton, Murphy, & Marlow, 1999). To improve the efficiency and eliminate this performance variation, some later works propose to use an adaptive threshold which can be dynamically determined based on video content (Lienhart, 1999; Dailianas et al., 1995). But even these methods require a lot of computational power in order to estimate successfully the appropriate threshold parameter, making their implementation a challenging issue, especially for real-time applications.

Thus, the research community faces the challenge of developing new techniques and methods for detecting scene changes over a video signal by eliminating the necessity of threshold parameters in the decision process.

Moreover, in order to allow a more efficient exploitation and browsing of video-anthologies, it is necessary to integrate these boundary shot detection techniques within content-based search and access methods, where the categorization of the video content is occurred by retrieving highly representative optical and semantic information. In this respect, the combination of frame extraction techniques and semantics will help towards the evolution of the current Web to the Semantic Web, where the browsing and searching of information will be based on semantic information.

CONCLUSION

This article outlines the various methods for detecting and extracting the scene changes from a video sequence. Depending on the metric that is exploited for the detection procedure, the methods that have been proposed are classified into two broad categories: those based on the uncompressed domain and those that exploit the metric of the compressed domain. Both the categories share the common drawback that they use threshold values for their decisions. Thus, the research community faces the challenge to develop new techniques that eliminate the use of threshold values, eliminating in this way the complexity and the computational requirements of the proposed methods.

ACKOWLEDGMENTS

This article is carried out within the "PYTHAGORAS II" research framework, jointly funded by the European Union and the Hellenic Ministry of Education.

REFERENCES

Dailianas, A., Allen, R. B., & England, P. (1995). Comparison of automatic video segmentation algorithms. *Proceedings of SPIE* (Vol. 2615, pp.2-16).

Hampapur, A., Jain, R., & Weymouth, T. E. (1995). Production model based digital video segmentation. *Multimedia Tools and Applications, 1*(1), 9-46.

Kasturi, R., & Jain, R. (1991). *Dynamic vision* (pp. 469-480). IEEE Computer Society Press.

Kikukawa, T., & Kawafuchi, S. (1992). Development of an automatic summary editing system for the audio visual resources. *Transactions on Electronics and Information,* 204-212.

Koprinska, I., & Carrato, S. (2001). Temporal video segmentation: A survey. *Signal Processing: Image Communication, 16,* 477-500.

Lienhart, R. (1999). Comparison of automatic shot boundary detection algorithms. *Proceedings of SPIE* (Vol. 3656, pp. 290-301).

Lu, H., & Tan, Y-P. (2005). An effective post-refinement method for shot boundary detection. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(11), 1407-1421.

Meng, J., Juan, Y., & Chang, S.-F. (1995). Scene change detection in a MPEG compressed video sequence. *Proceedings of the SPIE International Symposium on Electronic Imaging* (Vol. 2417, pp. 14-25), San Jose, CA.

O'Toole, C., Smeaton, A., Murphy, N., & Marlow, S. (1999). Evaluation of automatic shot boundary detection on a large video suite. *Proceedings of the 2nd UK Conference on Image Retrieval: The Challenge of Image Retrieval,* Newcastle, UK.

Shahraray, B. (1995). Scene change detection and content-based sampling of video sequences. *Proceedings of SPIE* (pp. 2-13).

Tam, W. J., Stelmach, L., Wang, L., Lauzon, D., & Gray, P. (1995, February 6-8). Visual masking at video scene cuts. *Proceedings of SPIE,* San Jose, CA.

Yeo, B., & Liu, B. (1995). Rapid scene analysis on compressed video. *IEEE Transactions on Circuits and Systems for Video Technology, 5*(6).

Zhang, H. J., Low, C. Y., Gong, Y. H., & Smoliar, S. W. (1994). Video parsing using compressed data. *Proceedings of the SPIE Conference of Image and Video Processing II* (pp.142-149).

Zhang, H. J., Low, C. Y., Gong, Y. H., & Smoliar, S. W. (1995). Video parsing and browsing using compressed data. *Multimedia Tools and Applications, 1,* 89-111.

KEY TERMS

Bit Rate: A data rate expressed in bits per second. In video encoding the bit rate can be *constant,* which means that it retains a specific value for the whole encoding process, or *variable,* which means that it fluctuates around a specific value according to the content of the video signal.

Frame: One of the many still images which as a sequence composes a video signal.

Histogram: A luminance- or color-level histogram of a frame depicts the density of the number of pixels that have specific luminance or color value.

Multimedia: The several different media types (e.g., text, audio, graphics, animation, video).

Pixel: Considered the smallest sample of a digital image or video.

Shot: An unbroken sequence of frames taken continuously from a single camera.

Video Codec: The device or software that enables the compression/decompression of digital video.

Video Coding: The process of compressing and decompressing a raw digital video sequence.