S

Shot Boundary Detection Techniques for Video Sequences

H. Koumaras
N.C.S.R., Demokritos, Greece

G. Xilouris
N.C.S.R., Demokritos, Greece

E. Pallis
Technological Educational Institute of Crete, Greece

G. Gardikis
University of the Aegean, Greece

A. Kourtis
N.C.S.R., Demokritos, 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:

$$Y = AXA^T$$

Related Content