Scene change detection techniques for video database systems

Haitao Jiang, Abdelsalam (Sumi) Helal, Ahmed K. Elmagarmid, Anupam Joshi

Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA; e-mail: {jiang, helal, ake, joshi}@cs.purdue.edu

Abstract. Scene change detection (SCD) is one of several fundamental problems in the design of a video database management system (VDBMS). It is the first step towards the automatic segmentation, annotation, and indexing of video data. SCD is also used in other aspects of VDBMS, e.g., hierarchical representation and efficient browsing of the video data. In this paper, we provide a taxonomy that classifies existing SCD algorithms into three categories: full-video-image-based, compressed-video-based, and model-based algorithms. The capabilities and limitations of the SCD algorithms are discussed in detail. The paper also proposes a set of criteria for measuring and comparing the performance of various SCD algorithms. We conclude by discussing some important research directions.

Key words: Scene change detection – Video segmentation – Video databases – Survey

1 Introduction

A video database management system is a software that manages a collection of video data and provides content–based access to users [10]. There are four basic problems that need to be addressed in a video database management system. These are video data modeling, video data insertion, video data storage organization and management, and video data retrieval. One fundamental aspect that has a great impact on all basic problems is the content–based temporal sampling of video data [24]. The purpose of the content–based temporal sampling is to identify significant video frames to achieve better representation, indexing, storage, and retrieval of the video data. Automatic content–based temporal sampling is very difficult due to the fact that the sampling criteria are not well defined, i.e., whether a video frame is important or not is usually subjective. Moreover, it is usually highly application–dependent and requires high–level, semantic interpretation of the video content. This requires the combination of very sophisticated techniques from computer vision and AI. The state of the art in those fields, however, has not advanced to the point where semantic interpretations would be possible.

However, researchers usually can get satisfying results by analyzing the visual content of the video and partitioning it into a set of basic units called shots. This process is also referred to as video data segmentation. Content–based sampling thus can be approximated by selecting one representing frame from each shot, since a shot is defined as a continuous sequence of video frames which have no significant inter–frame difference in terms of their visual contents.1 A single shot usually results from a single continuous camera operation. This partitioning is usually achieved by sequentially measuring inter–frame differences and studying their variances, e.g., detecting sharp peaks. This process is often called scene change detection (SCD).

Scene change in a video sequence can either be abrupt or gradual. Abrupt scene changes result from editing “cuts” (Fig. 1), and detecting them is called cut detection [11]. Gradual scene changes result from chromatic edits, spatial edits and combined edits [11]. Gradual scene changes include special effects like zoom, camera pan, dissolve and fade in/out, etc. An example of abrupt scene change and gradual scene change is shown in Fig. 2. SCD is usually based on some measurements of the image frame, which can be computed from the information contained in the images. This information can be color, spatial correlation, object shape, motion contained in the video image, or discrete cosine (DC) coefficients in the case of compressed video data. In general, gradual scene changes are more difficult to detect than abrupt scene changes, and may cause lots of scene detection algorithms to fail under certain circumstances.

Existing SCD algorithms can be classified in many ways according to, among others, the video features they use and the video objects they can be applied to. In this paper, we discuss SCD algorithms in three main categories: (1) approaches that work on uncompressed full–image sequences; (2) algorithms that aim at working directly on the compressed video; and (3) approaches that are based on explicit

1 There are many definitions in the literature from different points of views. This definition seems to be the one most agreed upon.
models. The latter are also called top-down approaches [10], whereas the first two categories are called bottom-up approaches. This paper is organized as follows. Section 2 briefly presents some background information about the SCD problem. Then, three categories of existing work are summarized in Sects. 3, 4, and 5, respectively. Their performance, advantages, and drawbacks are also discussed. Section 6 presents some criteria for evaluating the performance of SCD algorithms. Section 7 discusses some possible future research directions.

2 Background

We now introduce some basic notations used in this paper, followed by the notions of DC images, DC sequences and how they can be extracted from compressed video. Several most often used image measurements are also briefly described in terms of their use in measuring the inter-frame difference. It should be noted that they may not work well for scene detection when used separately, thus they usually are combined in the SCD algorithms. For example, Swanberg et al. [28] use a combination of template and histogram matching to measure the video frames.

2.1 Basic notations

Following notations are used throughout this paper. A sequence of video images, whether they are fully uncompressed or spatially reduced, are denoted as \( I_i, 0 \leq i < N \), \( N \) is the length or the number of frames of the video data. \( I_i(x, y) \) denotes the value of the pixel at position \((x, y)\) for the \(i\)th frame. \( H_i\) refers to the histogram of the image \( I_i\). The inter-frame difference between images \( I_i, I_j \) according to some measurement is represented as \( d(I_i, I_j) \).

2.2 MPEG standard: different frame types

According to the International Standard ISO/IEC 11172 [8], an MPEG-I compressed video stream can have one or more of the following types of frames:

- I (intra-coded) frames are coded without reference to other frames. They are coded using spatial redundancy reduction which is a lossy block–based coding involving DCT, quantization, run length encoding, and entropy coding.
- P (predicative-coded) frames are coded using motion compensation predication from the last I or P frame.
- B (bidirectionally predictive coded) frames are coded using motion compensation with reference to both previous and next I or P frames.
- D (DC-coded) frames are coded using DC coefficients of blocks, thus only contain low-frequency information. D frames are not allowed to co-exist with I/P/B frames and are rarely used in practice.

Obviously, any MPEG compressed video stream must have at least I frames. The data size ratios between frames suggested by the standard are: 3:1 for I:P and 5:2 to 2:1 for P:B. In other words, B frames have the highest degree of compression and I frames have the least one. More details about MPEG video streams can be found in [8].

2.3 DC images, DC sequences and their extraction

A DC image [31–34] is a spatially reduced version of a given image. It can be obtained by first dividing the original image into blocks of \( n \times n \) pixels each, then computing the average value of pixels in each block, which corresponds to one pixel in the DC image. For the compressed video data, e.g., MPEG video, a sequence of DC images can be constructed directly from the compressed video sequence, which is called a DC sequence. Figure 3 is an example of a video frame image and its DC image.

There are several advantages of using DC images and DC sequences in the SCD for the compressed video.