Visual Information Analysis for Interactive TV Applications

Evlampios Apostolidis
Information Technologies Institute, Centre for Research and Technology Hellas, Greece

Panagiotis Sidiropoulos
Information Technologies Institute, Centre for Research and Technology Hellas, Greece

Vasileios Mezaris
Information Technologies Institute, Centre for Research and Technology Hellas, Greece

Ioannis Kompatsiaris
Information Technologies Institute, Centre for Research and Technology Hellas, Greece

INTRODUCTION

Since its introduction, television has been providing to millions of users a non-interactive experience, in which viewers can only participate as passive consumers of audiovisual content. Recently, the extreme proliferation and success of the Internet and the widespread appraisal of the interaction possibilities that it offers gave rise to the idea of the Interactive TV: a television broadcast in which users do not only passively consume the content, but similarly to the Web, they can navigate across multiple pieces of content, following links that are similar in nature to the hypertext links between textual documents.

In this article, we will discuss the visual information analysis technologies and tools that are necessary for supporting the interlinking of visual content in a fashion that allows users to navigate between fragments of the content. We will cover analysis technologies that range from video content fragmentation into temporal units (shots, scenes), to the labeling of visual content via concept-based annotation and re-detection of objects of interest in the video. Such technologies are necessary for empowering video hyperlinking, so that e.g. an object of interest in one video segment can be linked to other relevant segments of the same video, or also to entirely different videos that relate to it. For these key-enabling analysis technologies, we will review the state-of-the-art and we will further elaborate on and provide indicative results for specific techniques that are particularly relevant to TV content.

DOI: 10.4018/978-1-4666-5888-2.ch214

BACKGROUND

Temporal Video Segmentation

Temporal video segmentation aims to partition the video into elementary temporal units. These are typically either shots, where each shot is defined as a set of consecutive frames taken without interruption by a single camera, or scenes, which are the basic story-telling units of the video and may consist of one or more shots.

Shot Segmentation

Segmentation to shots is performed by detecting the transition from one shot to the next. Transitions between shots can be abrupt and gradual, where in the latter case the visual content of two consecutive shots is combined using video production effects, such as fade in/out, wipe and dissolve (see Figure 1).

The simplest technique for shot segmentation in uncompressed video is based on pair-wise pixel comparisons between successive or distant frames. However, such methods are sensitive to small camera and object motion, which is why many researchers proposed the use of color histograms. Histogram-based methods rely on the comparison or the intersection of color histograms from successive frames, calculated either at the image or at a more detailed block level (Tan, Teng, & Zhang, 2007). Alternatively, a more

Figure 1. Examples of abrupt and gradual transitions between shots

(a) Abrupt transition. The last frame of the first shot (left image) is immediately followed by the first frame of the next shot (right image).

(b) Gradual transition. The last frame of the first shot (top left) is gradually replaced by the first frame of the second shot (bottom right).

sophisticated algorithm that combines color histograms with luminance information and performs shot boundary detection using Support Vector Machine (SVM) classifiers was presented in (Tsamoura, Mezaris, & Kompatsiaris, 2008). Following the development of scale- and rotation-invariant local descriptors (e.g. SIFT (Lowe, 2004)) some temporal segmentation approaches that are based on them were also proposed. For example, Lankinen and Kämäräinen (2013) introduced a method that is based on detecting and tracking objects between successive frames of the video. For this purpose, they extract SIFT descriptors from the video frames and cluster them into a predefined number of bins, creating a codebook of visual words. Each frame then is expressed by a histogram of words and the shot boundaries are determined based on the similarity of the formed histograms.

Other efforts on shot boundary detection avoid the prior decompression of the video stream, resulting to significant gains in efficiency. Such methods consider mostly MPEG video and exploit compression-specific cues to detect points in the 1D (i.e., temporal) decision space where redundancy, which is inherent in video and is greatly exploited by compression schemes, is reduced. These cues can be macro-block type information of specific frames and DC coefficients or motion vectors that are included in the compressed data stream (Doulaverakis, Vagionitis, Zervakis, & Petrakis, 2004). Regardless of whether the temporal segmentation is applied to raw or compressed video, it is often accompanied by the selection of one or more representative key-frames per shot, which can be used for further processing (e.g. concept detection). The selection process can be as simple as selecting by default the first or median frame of the shot, or can be more elaborate, as in (Liu & Fan, 2005), where the algorithm clusters the frames of a shot into a predefined number of bins (equal to the desired number of key-frames) based-on frame-wise histogram comparisons.

Scene Segmentation

Several approaches were proposed for decomposing video into scenes. Mainly they group shots into scenes using similarity measures and temporal consistency rules. For estimating the similarity between shots (most often approximated by comparing their representative key-frames), various local or global characteristics can be employed, such as edges, color, texture and motion.