Semantic Video Analysis and Understanding

Vasileios Mezaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Georgios Th. Papadopoulos
Aristotle University of Thessaloniki, Greece
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Alexia Briassouli
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Ioannis Kompatsiaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Michael G. Strintzis
Aristotle University of Thessaloniki, Greece
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

INTRODUCTION

Access to video content, either amateur or professional, is nowadays a key element in business environments, as well as everyday practice for individuals all over the world. The widespread availability of inexpensive video capturing devices, the significant proliferation of broadband Internet connections and the development of innovative video sharing services over the World Wide Web have contributed the most to the establishment of digital video as a necessary part of our lives. However, these developments have also inevitably resulted in a tremendous increase in the amount of video material created every day. This presents new possibilities for businesses and individuals alike. Business opportunities in particular include the development of applications for semantics-based retrieval of video content from the Internet, video stock agencies or personal collections; semantics-aware delivery of video content in desktop and mobile devices; and semantics-based video coding and transmission. Evidently, the above opportunities also reflect to the video manipulation possibilities offered to individual users. Besides opportunities, though, the abundance of digital video content also presents new and important technological challenges, which are crucial for the further development of the aforementioned innovative services.

The cornerstone of the efficient manipulation of video material is the understanding of its underlying semantics, a goal that has long been identified as the "Holy grail of content-based media analysis research" (Chang, 2002). Efforts to understand the semantics of video content typically build on algorithms that operate at the signal level, such as temporal and spatiotemporal video segmentation algorithms that aim at partitioning a video stream into semantically meaningful parts. To support the goal of semantic analysis, these signal-level algorithms are augmented with a priori knowledge regarding the different semantic objects and events of interest that may appear in the video and their signal-level properties. The introduction of a priori knowledge serves the purpose of facilitating the detection and exploitation of the hidden associations between the signal and semantic levels, resulting in the generation of semantically meaningful metadata for the video content.

In this article, existing state-of-the-art semantic video analysis and understanding techniques are reviewed, including a hybrid approach to semantic video analysis that is outlined in some more detail, and the future trends in this research area are identified. The literature presentation starts in the following section with signal level algorithms for processing video content, a necessary prerequisite for the subsequent application of knowledge-based techniques.

BACKGROUND

Segmentation is in general the process of partitioning a piece of information into meaningful elementary parts termed segments. Considering video, the term segmentation is used to describe a range of different processes for partitioning the video into meaningful parts at different granularities (Salembier & Marques, 1999). Segmentation of video can thus be temporal, aiming to break down the video to scenes or shots, spatial, addressing the problem of independently

segmenting each video frame to arbitrarily shaped regions, or spatio-temporal, extending the previous case to the generation of temporal sequences of arbitrarily shaped spatial regions. The term segmentation is also frequently used to describe foreground/background separation in video, which can be seen as a special case of spatio-temporal segmentation. In any case, the application of any segmentation method is often preceded by a simplification step for discarding unnecessary information (e.g., low-pass filtering) and a feature extraction step for modifying or estimating features not readily available in the visual medium (e.g., texture, motion features , etc., but also color features in a different color space, etc.).

Temporal Video Segmentation

Temporal video segmentation aims to partition the video to elementary image sequences termed shots. A shot is defined as a set of consecutive frames taken without interruption by a single camera. A scene, on the other hand, is usually defined as the basic story-telling unit of the video, that is, as a temporal segment that is elementary in terms of semantic content and may consist of one or more shots.

Temporal segmentation to shots is performed by detecting the transition from one shot to the next. Transitions between shots, which are effects generated at the video editing stage, may be abrupt or gradual, the former being detectible by examining two consecutive frames, the latter spanning more than two frames and being usually more difficult to detect, depending among others on the actual transition type (e.g., fade, dissolve, wipe, etc.). Temporal segmentation to shots in uncompressed video is often performed by means of pair-wise pixel comparisons between successive or distant frames or by comparing the color histograms corresponding to different frames. Methods for histogram comparison include the comparison of absolute differences between corresponding bins and histogram intersection (Gargi, Kasturi, & Strayer, 2000). Other approaches to temporal segmentation include block-wise comparisons, where the statistics of corresponding blocks in different frames are compared and the number of "changed" blocks is evaluated by means of thresholding, edge-based and motion-based methods.

Other recent efforts on shot detection have focused on avoiding the prior decompression of the video stream, resulting to significant gains in terms of efficiency. Such methods consider mostly MPEG video, but also other compression schemes such as wavelet-based ones. These exploit compression-specific cues such as macroblock-type ratios to detect points in the 1D decision space where temporal redundancy, which is inherent in video and greatly exploited by compression schemes, is reduced. Regardless of whether the temporal segmentation is applied to raw or compressed video, it is often accompanied by a procedure for selecting one or more representative key-frames of the shot; this can be as simple as selecting by default the first or median frame of the shot or can be more elaborate, as for example in Liu and Fan (2005), where a combined key-frame extraction and object segmentation approach is proposed.

Spatial and Spatio-Temporal Segmentation

Several approaches have been proposed for spatial and spatio-temporal video segmentation (i.e., segmentation in a 2D and 3D decision space, respectively), both unsupervised and supervised. The latter require human interaction for defining the number of objects present in the sequence, for estimating an initial contour of the objects to be tracked or for grouping homogeneous regions to semantic objects, while the former require no such interaction. In both types of approaches, it is typically assumed that spatial or spatio-temporal segmentation is preceded by temporal segmentation to shots and possibly the extraction of one or more key-frames, as discussed in the previous section.

Segmentation methods for 2D images may be divided primarily into region-based and boundary-based methods. Region-based approaches rely on the homogeneity of spatially localized features such as intensity, texture, and position. They include among others the K-means algorithm and evolved variants of it, such as K-Means-with-Connectivity-Constraint (Mezaris, Kompatsiaris, & Strintzis, 2004a), the Expectation-Maximization (EM) algorithm (Carson, Belongie, Greenspan, & Malik, 2002), and Normalized Cut, which treats image segmentation as a graph partitioning problem (Shi & Malik, 2000). Boundary-based approaches, on the other hand, use primarily gradient information to locate object boundaries. They include methods such as anisotropic diffusion, which can be seen as a robust procedure for estimating a piecewise smooth image from a noisy input image (Perona & Malik, 1990). Additional techniques for spatial segmentation include mathematical morphology methods, in particular the watershed algorithm, and global energy minimization schemes, also known as snakes or active contour models.

Regarding spatio-temporal segmentation approaches, some of them rely on initially applying spatial segmentation to each frame independently. Spatio-temporal objects are subsequently formed by associating the spatial regions formed in successive frames using their low-level features (Deng & Manjunath, 2001). A different approach is to use motion information to perform motion projection, that is, to estimate the position of a region at a future frame, based on its current position and its estimated motion features. In this case, a spatial segmentation method need only be applied to the first frame of the sequence, whereas in subsequent frames only refinement of the motion projection result is required (Tsai, Lai, Hunga, & Shih, 2005). A similar approach is followed in Mezaris, Kompatsiaris, and Strintzis (2004b), where the need for motion projection is substituted by a Bayes-based approach to color-homogeneous region-tracking