Video Event Understanding

D

Nikolaos Gkalelis
Information Technologies Institute, Centre for Research and Technology Hellas, Greece

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

Michail Dimopoulos
Information Technologies Institute, Centre for Research and Technology Hellas, Greece

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

INTRODUCTION

The automatic understanding of multimedia content, particularly capturing the meaning that real-life image and video content conveys, has long been one of the major challenges in the multimedia community. Most recently, triggered by the realization that the human understanding and recollection of visual information revolves around the notion of event, the detection and understanding of high-level events that are depicted on or otherwise relate to the video content, rather than merely physical object detection or broad content categorization, has started to receive significant attention. Automatic video event detection and understanding is now widely recognized as an essential step towards large-scale multimedia content analysis, indexing and search.

In this article, after introducing the notion of video event and providing motivation for the event-based analysis and organization of video content, we will briefly review the existing literature on complex video event detection. Subsequently, we will introduce in some more detail a very promising class of techniques that use traditional concept or category-based video analysis results as the stepping stone for detecting complex events in video. Indicative results and additional references to individual technologies that contribute to this class of techniques will provide the reader with a comprehensive overview of the state of the art in the timely research topic of video event understanding.

BACKGROUND

High-level events can be conceived as dynamic objects that pace our everyday activities and index our memories. This definition reflects the compositional nature of the event (i.e., consisting of actions, actors, objects, locations, times and other components with possible relations among them), and implies that its perception depends on the cultural and personal perspective of the observer (Brown, 2005). For this reason, it is generally expected that event understanding technologies can offer effective organization of multimedia content and natural language description of this content to human users. On the other hand, it is clear that this task is much more challenging than tasks dealing with the detection of simple human actions (Turaga, Chellappa, Subrahmanian, & Udrea, 2008) or other semantic concepts (Mezaris, Papadopoulos, Briassouli, Kompatsiaris, & Strintzis, 2008).

The necessity of event models for describing real life events in video signals has been recently acknowledged as an essential step towards effective large-scale multimedia content analysis, indexing and search (Gupta & Jain, 2011). Moreover, in Westermann and Jain (2007) a set of aspects that an event model should satisfy are defined, such as media independence, model interoperability, and other. In the following we briefly review a representative fraction of the related work. In Scherp, Franz, Saathoff, and Staab (2009), the pattern oriented ontology approach of DOLCE and DUL is utilized to define the event-model F so that several event aspects (Westermann & Jain, 2007) are addressed. In Gupta

DOI: 10.4018/978-1-4666-5888-2.ch213

and Jain (2011), the event-model E* is presented that extends the event-model E (Westermann & Jain, 2007) using a graph-based design and the ABC and DOLCE ontology to provide formal definition of event aspects. In Gkalelis, Mezaris, and Kompatsiaris (2010a, 2010b) a joint content-event model is presented, which additionally provides a mechanism for the automatic (or semi-automatic) association and enrichment of event descriptions with multimedia content, so that video event analysis technologies can be directly exploited for populating the model. Combining event models and video annotation tools, Agius and Angelides (2006) proposed COSMOS-7, an MPEG-7 compliant scheme for modeling events, objects and spatiotemporal relationships among them, and based on it they designed COSMOSIS to enable annotation of video content. Several other video annotation tools that support the generation of event-based video descriptions have been presented, such as Vannotator (http://www.vannotator.com), CAVIAR (http://homepages.inf.ed.ac.uk/rbf/CAVIAR/), and other.

For the automatic instantiation and enrichment of the event models described above, efficient and effective event analysis algorithms are required. During the past few years there has been a surge of research in the area of high-level event detection in video signals (Jiang, Bhattacharya, Chang, & Shah, 2012). Event detection algorithms typically derive one or more low-level features and combine the different feature modalities using a fusion strategy. The different fusion possibilities include: a) train one classifier for each modality and combine their outputs (late fusion), b) combine the low-level features and train a single classifier (early fusion), and, c) exploit a framework that utilizes both late and early fusion. For instance, in Kamishima et al. (2011), a variety of features (Harris-SIFT, Hessian-SIFT, space time interest points-HOG (STIP-HOG), STIP-HOF, dense HOG, MFCC) are extracted, and a Gaussian mixture model (GMM) supervector is constructed for each feature and each video. The derived GMM supervectors are used to train one kernel SVM (KSVM) for each event in the TRECVID 2012 MED dataset, and the weighted average of the KSVM output scores is exploited for event detection.

Recently, a few researchers exploit a semantic model vector as a feature representation of high-level events, aiming at better event detection performance. The inspiration behind this modeling approach is that high-level events can be better recognized by looking at their constituting semantic entities. For instance, in (Merler, Huang, Xie, Hua, & Natsev, 2012) a set of pre-trained concept detectors are used for describing the video signal. In Gkalelis, Mezaris, & Kompatsiaris (2011b) and Moumtzidou, Dimou, Gkalelis, Vrochidis, Mezaris, and Kompatsiaris (2010) discriminant analysis is additionally used to derive the most informative event concepts, which are then used for describing the videos and for learning the target events. Experimental results in the above works showed that in some cases event detectors trained using the semantic model vector representation outperformed classifiers trained on state-of-the-art low-level feature representations alone, and that their combination with low-level features provides small but noticeable performance gains. However, the main advantage of using a semantic model vector approach is that the derived video representation can serve as the first step towards automatic concept- and event-based textual description of the videos (Gkalelis, Mezaris, & Kompatsiaris, 2010a, 2010b). On that basis, it is anticipated that this type of approaches are going to have a significant contribution in the emerging field of video event understanding. For instance, the most informative semantic concepts can be used to enumerate the key semantic evidences of the event in multimedia event recounting (MER) tasks (Yu, Liu, Cheng, Divakaran, & Sawhney, 2012), or for leveraging knowledge in the Ad Hoc multimedia event detection (MED) task (Ma, Yang, Cai, Sebe, & Hauptmann, 2012). For the above reasons, in the next section we provide a comprehensive treatment of selected event detection approaches utilizing a model vector-based video representation.

EVENT DETECTION TECHNIQUES

In this section we present two recently developed model vector-based approaches for event detection. The first method utilizes the most discriminant semantic concepts concerning the target event (Gkalelis, Mezaris, & Kompatsiaris, 2011b; Moumtzidou, Dimou, Gkalelis, Vrochidis, Mezaris, & Kompatsiaris, 2010; Tsampoulatidis, Gkalelis, Dimou, Mezaris, & Kompatsiaris, 2011), while the second exploits a subclass error correcting output (SECOC) framework to combine multiple event detectors (Gkalelis, Mezaris, Kompatsiaris, & Stathaki,