Encyclopedia of Information Science and Technology

Second Edition

Mehdi Khosrow-Pour
Information Resources Management Association, USA

Volume VII
R–S

All work contributed to this encyclopedia set is original material. The views expressed in this encyclopedia set are those of the authors, but not necessarily of the publisher.

Semantic Video Analysis and Understanding

S

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

using color information, and the resulting spatio-temporal regions are eventually clustered to different objects using their long-term motion trajectories.

Alternatively to the above techniques, one could restrict the problem of video segmentation to foreground/background separation. In Chien, Huang, Hsieh, Ma, and Chen (2004), a fast moving object segmentation algorithm is developed, based upon change detection and background registration techniques; this algorithm also incorporates a shadow cancellation technique for dealing with light changing and shadow effects.

Finally, as in temporal segmentation, the spatio-temporal segmentation of compressed video has recently attracted considerable attention. Algorithms of this category generally employ coarse motion and color information that can be extracted from the video stream without full decompression, such as macroblock motion vectors and DC coefficients of DCT-coded image blocks (Mezaris, Kompatsiaris, Boulgouris, & Strintzis, 2004).

SEMANTIC VIDEO ANALYSIS AND UNDERSTANDING TECHNIQUES

The result of pure segmentation techniques, though conveying some semantics, such as the complexity of the key-frame or video, measured by the number of generated regions, or the existence of moving objects in the shot, is still far from revealing the complete semantic content of the video. To alleviate this problem, the introduction of prior knowledge to the segmentation procedure, leading to the development of domain-specific knowledge-assisted analysis techniques, has been proposed.

Prior knowledge for a domain (e.g., F1 racing) typically includes the important objects that can be found in any given image or frame belonging to this domain (e.g., car, road, grass, sand, etc.), their characteristics (e.g., corresponding color models) and any relations between them. Given this knowledge, there exists the well-posed problem of deciding, for each pixel, whether it belongs to any of the defined objects (and if so, to which one) or to none of them.

Depending on the adopted knowledge acquisition and representation process, two types of approaches can be identified in the relevant literature: implicit, realized by machine learning methods, and explicit, realized by model-based approaches. The usage of machine learning techniques has proven to be a robust methodology for discovering complex relationships and interdependencies between numerical image data and the perceptually higher-level concepts. Moreover, these elegantly handle problems of high dimensionality. Among the most commonly adopted machine learning techniques are Neural Networks (NNs), Hidden Markov Models (HMMs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Genetic Algorithms

(GAs) (Assfalg, Berlini, Del Bimbo, Nunziat, & Pala, 2005; Zhang, Lin, & Zhang, 2001). On the other hand, model-based video analysis approaches make use of prior knowledge in the form of explicitly defined facts, models and rules, that is, they provide a coherent semantic domain model to support "visual" inference in the specified context (Hollink, Little, & Hunter, 2005).

Knowledge Representation and Ontologies

Ontology, being a formal specification of a shared conceptualization (Gruber, 1993), provides by definition the formal framework required for exchanging interoperable knowledge components. By making semantics explicit to machines, ontologies enable automatic inference support, thus allowing users, agents, and applications to communicate and negotiate over the meaning of information. Typically, an ontology identifies classes of objects that are important for the examined subject area (domain) under a specific viewpoint and organizes these classes in a taxonomic (i.e., subclass/super-class) hierarchy. Each such class is characterized by properties that all elements (instances) in that class share. Important relations between classes or instances of the classes are also part of the ontology. Consequently, ontologies can be suitable for expressing multimedia content semantics so that automatic semantic analysis and further processing of the extracted semantic descriptions is allowed (Hollink et al., 2005). In (Dasiopoulou, Mezaris, Papastathis, Kompatsiaris, & Strintzis, 2005), an ontology was developed for representing the knowledge components that need to be explicitly defined for video analysis. More specifically, domain knowledge is combined with object low-level features and spatial descriptions realizing an ontology- aided video analysis framework. To accomplish this, F-logic rules (Angele & Lausen, 2004) are used to relate the extraction of the semantic concepts, the execution order of the necessary multimedia processing algorithms and the low-level features associated with each semantic concept, thus integrating knowledge and intelligence in the analysis process. The overall system thus consists of a knowledge base, an inference engine, the algorithm repository containing the necessary multimedia analysis tools and the system main processing module, which performs the analysis task, using the appropriate sets of tools and multimedia features, for the semantic multimedia description extraction.

Semantic Video Analysis Approaches

Few semantic video (or, visual content in general) analysis approaches have been presented so far in the literature. Starting with image analysis, a knowledge-guided segmentation and labeling approach for still images is presented in Zhang,

Hall and Goldgof (2002), where an unsupervised fuzzy C-means clustering algorithm along with basic image processing techniques are used under the guidance of a knowledge base. The latter is constructed by automatically processing a set of ground-truth images to extract cluster-labeling rules.

In Naphade, Kozintsev and Huang (2002), the understanding of the semantics of the video content for the purpose of indexing and in particular the association of low-level representations and high-level semantics is formulated as a probabilistic pattern recognition problem and is addressed with the introduction of a factor graph framework. Another approach to video semantic object detection is presented in Tsechpenakis, Akrivas, Andreou, Stamou and Kollias (2002), where semantic entities in the context of the MPEG-7 standard are defined; moving regions are extracted at the signal level by an active contour technique and then their low-level features are matched against those assigned to the previously defined semantic entities, resulting in the identification of associations between the latter and the video segments corresponding to moving objects. In Dasiopoulou et al. (2005), domain knowledge is combined with object low-level features and spatial descriptions realizing an ontology-aided video analysis framework, as described in the previous section. This approach is extended in Voisine et al. (2005), where a genetic algorithm is applied to the atom regions initially generated via simple segmentation in order to find the optimal scene interpretation according to the domain conceptualization. With respect to event detection, existing approaches include among others Sadlier and O'Connor (2005), where a framework for event detection in broadcast video of multiple different field sports is developed, based on robust detectors.

A Hybrid Approach to Semantic Video Analysis

In this section, an example of a semantic video analysis approach that combines two types of machine learning algorithms, namely SVMs and GAs, with explicitly defined domain-specific knowledge in the form of an ontology, is discussed Papadopoulos, Panagi, Dasiopoulou, Mezaris, and Kompatsiaris (2006). SVMs are used for acquiring the implicit knowledge that is required for the analysis process. Additionally, an ontology is developed for representing the knowledge components that need to be explicitly defined, that is, the semantic objects of interest for the selected domain, as well as their spatial relations. The latter appear in the form of fuzzy directional relations, which are used for denoting the relative positions of the depicted real-world objects. The GA is employed for exploiting the aforementioned spatial-related contextual information and deciding upon the final image semantic interpretation.

According to this approach, the video sequence is initially segmented into shots (temporal segmentation) and key-frames

Figure 1. Outline of hybrid semantic video analysis approach
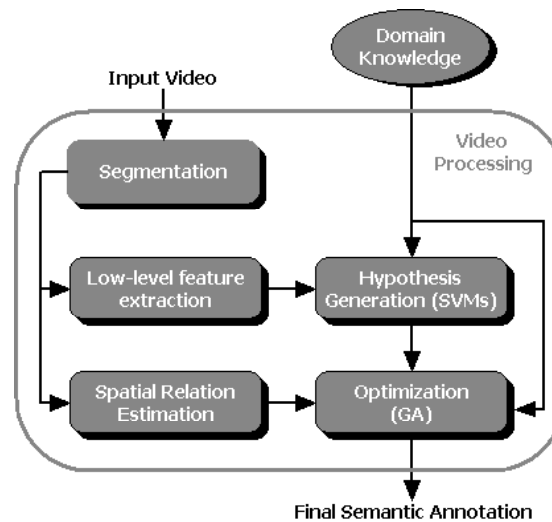
are extracted. Then, for every resulting key-frame, spatial segmentation is performed and, subsequently, low-level visual features and fuzzy directional relations are estimated at the region level. Region low-level visual features include the Scalable Color, Homogeneous Texture, Region Shape and Edge Histogram standardized MPEG-7 descriptors. Following the extraction of low-level visual features, SVMs employ them for performing an initial mapping between the image regions and the domain objects in the developed ontology (i.e., generating an initial hypothesis set for every image region). Finally, after the application of SVMs to each region independently, a GA is used to optimize the region-object mapping over the entire image, taking into account the region fuzzy directional relations and the corresponding spatial-related contextual knowledge that is stored in the ontology. This architecture is schematically presented in Figure 1. Application of the proposed approach to video frames of the specified domain results in the generation of fine granularity semantic annotations, that is, segmentation maps with semantic labels attached to each segment. A sample analysis outcome of this hybrid approach is illustrated in Figure 2.

FUTURE TRENDS

In the future, efforts will continue to concentrate on two different subfields of semantic video analysis and understanding: (a) the low-level analysis aiming at effectively decomposing the signal into semantically meaningful entities or parts thereof, without concentrating on understanding

Figure 2. Sample analysis outcome of a hybrid approach to semantic video analysis

Original Frame

Semantic Annotations

the actual semantics of them, and (b) the efficient use of knowledge, including the resolution of issues regarding the efficient acquisition, formulation and representation of it, for understanding the semantics of the signal parts formed after the application of a low-level analysis method. Of particular attention will most certainly be the closer interaction between the two aforementioned subfields: low level image analysis can benefit from the use of some form of knowledge about the content, though this may have to be represented in, for example, the form of a probability distribution (Freedman & Zhang, 2004), and thus may be different in nature from the knowledge used for understanding the semantics of already formed parts of the visual medium. Similarly, the latter task will most probably require the extension of existing knowledge representation formalisms, which were originally designed for tasks far different from the semantic analysis of video or multimedia content in general, so as to address the needs of analysis and provide support for improved reasoning based on the output of the latter.

CONCLUSION

In this article, semantic video analysis and understanding was discussed, starting with a literature review of the elementary task of video segmentation, which constitutes the first necessary step for the analysis and understanding of video content. Following that, the article focused on techniques that go beyond traditional segmentation to incorporate prior knowledge in the analysis procedure, so as to extract a high-level representation of the video content comprising semantic class memberships and recognized video objects. The dominant types of approaches presented in the literature for accommodating this task were identi-

fied and a hybrid approach combining machine learning algorithms with explicitly defined knowledge in the form of an ontology was presented in more detail. The future trends identified in the relevant section provide insights on how the algorithms outlined or presented in more detail in this article can be further evolved, so as to more efficiently address the problem of semantic video analysis and understanding and consequently pave the way for the development of innovative video applications.

REFERENCES

Angele, J., & Lausen, G. (2004). Ontologies in f-logic. *International handbooks on information systems.* Berlin, Germany.

Assfalg, J., Berlini, M., Del Bimbo, A., Nunziat, W., & Pala, P. (2005). Soccer highlights detection and recognition using HMMs. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME),* (pp. 825-828).

Carson, C., Belongie, S., Greenspan, H., & Malik, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(8), 1026-1038.

Chang, S. F. (2002). The holy grail of content-based media analysis. *IEEE Multimedia*, *9*(2), 6-10.

Chien, S. Y., Huang, Y. W., Hsieh, B. Y., Ma, S. Y., & Chen, L. G. (2004). Fast video segmentation algorithm with shadow cancellation, global motion compensation, and adaptive threshold techniques. *IEEE Transactions on Multimedia, 6*(5), 732-748.

Dasiopoulou, S., Mezaris, V., Papastathis, V. K., Kompatsiaris, I., & Strintzis, M. G. (2005). Knowledge-assisted semantic video object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(10), 1210-1224.

Deng, Y., & Manjunath, B.S. (2001). Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(8), 800-810.

Freedman, D., & Zhang, T. (2004). Active contours for tracking distributions. *IEEE Transactions on Image Processing, 13*(4), 518-526.

Gargi, U., Kasturi R., & Strayer, S. H. (2000). Performance characterization of video-shot-change detection methods. *IEEE Transaction on Circuits and Systems for Video Technology, 10*(1), 1-13.

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199-220.

Hollink, L., Little, S., & Hunter, J. (2005). Evaluating the application of semantic inferencing rules to image annotation. In Proceedings of the *3rd International Conference on Knowledge Capture (K-CAP05),* Banff, Canada.

Liu, L., & Fan, G. (2005). Combined key-frame extraction and object-based video segmentation. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(7), 869-884.

Mezaris, V., Kompatsiaris, I., Boulgouris, N. V., & Strintzis, M. G. (2004). Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(5), 606-621.

Mezaris, V., Kompatsiaris I., & Strintzis, M. G. (2004a). Still image segmentation tools for object-based multimedia applications. *International Journal of Pattern Recognition and Artificial Intelligence, 18*(4), 701-725.

Mezaris, V., Kompatsiaris, I., & Strintzis, M. G. (2004b). Video object segmentation using bayes-based temporal tracking and trajectory-based region merging. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(6), 782-795.

Naphade, M. R., Kozintsev, I.V., & Huang, T.S. (2002). A factor graph framework for semantic video indexing. *IEEE Transactions on Circuits and Systems for Video Technology, 12*(1), 40-52.

Papadopoulos, G. T., Panagi, P., Dasiopoulou, S., Mezaris, V., & Kompatsiaris, I. (2006, September). A learning approach to semantic image analysis. In *Proceedings of the 2nd International Mobile Multimedia Communications Conference (Mobimedia),* Alghero, Sardinia, Italy.

Perona, P., & Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(7), 629-639.

Sadlier, D. A., & O'Connor, N. E. (2005). Event detection in field sports video using audio-visual features and a support vector Machine. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(10), 1225-1233.

Salembier, P., & Marques, F. (1999). Region-based representations of image and video: Segmentation tools for multimedia services. *IEEE Transactions on Circuits and Systems for Video Technology, 9*(8), 1147-1169.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), 888-905.

Tsai, Y. P., Lai, C. C., Hunga, Y. P., & Shih, Z. C. (2005). A bayesian approach to video object segmentation via merging 3-D watershed volumes. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(1), 175-180.

Tsechpenakis, G., Akrivas, G., Andreou, G., Stamou, G., & Kollias, S.D. (2002, September). Knowledge-assisted video analysis and object detection. In *Proceedings of the European Symposium on Intelligent Technologies, Hybrid Systems and Their Implementation on Smart Adaptive Systems (Eunite02)*, Algarve, Portugal.

Voisine, N., Dasiopoulou, S., Mezaris, V., Spyrou, E., Athanasiadis, T., Kompatsiaris, I., et al. (2005, April). Knowledge-assisted video analysis using a genetic algorithm. In *Proceedings of the Workshop on Image Analysis For Multimedia Interactive Services,* Montreux, Switzerland.

Zhang, M., Hall, L. O., & Goldgof, D. B. (2002). A generic knowledge-guided image segmentation and labeling system using fuzzy clustering algorithms. *IEEE Transactions on Systems, Man and Cybernetics, Part B, 32*(5), 571-582.

Zhang, L., Lin, F.Z., & Zhang, B. (2001, October). Support vector machine learning for image retrieval. In *Proceedings of the International Conference on Image Processing.*

KEY TERMS

Compressed Video Segmentation: Segmentation of video without its prior decompression.

Knowledge-Assisted Analysis: Analysis techniques making use of prior knowledge for the content being processed.

Machine Learning Techniques: Training-based techniques for discovering and representing implicit knowledge, such as complex relationships and interdependencies between numerical image data and perceptually higher-level concepts.

Ontology: Knowledge representation formalism, used for expressing explicit knowledge.

Semantic Video Analysis: Extraction of the semantics of the video, that is, detection and recognition of semantic objects and events.

Spatiotemporal Video Segmentation: Partition the video to elementary spatio-temporal objects, that is, sequences of temporally adjacent arbitrarily-shaped spatial regions.

Temporal Video Segmentation: Partition the video to elementary image sequences termed shots, defined as a set of consecutive frames taken without interruption by a single camera.