Multimedia Tools and Applications, 26, 255–257, 2005

Guest Editorial: Special Issue on Video Segmentation for Semantic Annotation and Transcoding

Understanding of the semantic content of digital documents has been one of the ultimate goals of information and communication technology in the recent years. This allows effective search from large distributed databases where content feeding and annotation is out of the control of the final user. *Automatic semantic annotation* is one of the mayor requirements for content-based search from non-conventional databases where annotation is otherwise error-prone and too much time consuming to achieve manually. It is particularly challenging for video, due to the diversity of content and requirements among the different types of video, and also to the fact that, in images, events and entities are often cluttered and appear in manifold aspects inter-related each other. For example, in news videos, the anchorperson's subject of speech, or the reporter's service can be extracted so as to provide indexes for content-based search; in sports videos, the most relevant highlights of the game can be identified so as to create summaries of the key actions; in movies, music and audio streams, melody or pitch identification can help to discover parts that have the same mood or keep the distinguishing elements as a note of the overall document.

Automatic video annotation is also important for *semantic video transcoding*. Video transcoding concerns the changing of the video format with respect to color, size, resolution, code type, etc., in relation with the channel constraints and the client requirements. It is particularly important whenever smart phones, PDAs, mobile terminals must access multimedia data with low bandwidth channels. If transcoding is merely a syntactic procedure, video changes are based only on communication constraints, without considering video content and user's interests. Semantic video transcoding, instead, aims at changing the format of video according to the video content, the communication constraints and the user requirements. In this framework, semantic annotation is mandatory. By detecting the most significant parts of the video, semantic annotation permits that data are transmitted at different formats and qualities depending on what content is represented, its relevance and correspondence with the users' preferences. For example, users accessing sports videos might be interested in players and attack actions. Users accessing news videos might be instead interested in audio more than in than visual information, so that simple *transmoding* from video to audio can be acceptable. Actually, semantic transcoding can be regarded not simply as a way for changing video coding but also as a way for providing *video abstraction* and *video summarization,* by clustering and saving significant temporal segments of the video.

In this Special Issue we have collected several papers that highlight distinct aspects of the relationships between semantic annotation and transcoding of videos. They

address segmentation of video into shots, highlights and objects of interests and propose different methods of endowed knowledge representation, together with models to manage and access video entities efficiently and methodologies to evaluate results and performances.

The paper "*Ontology-based semantic indexing for MPEG-7 and TV-Anytime audiovisual content*" by C. Tsinaraki et al. provides a complete presentation of different methodologies and tools for video segmentation and annotation and for ontology-based indexing and retrieval. The approach is compliant with the MPEG-7 and TV-Anytime standard specifications for metadata description. A complete ontology for soccer games is defined and used as a case study.

The paper by N.Petrovic et al., "*Adaptive Video Fast Forward*" addresses instead the subject of non-uniform access to video content based on semantic annotation. It exploits the semantics extracted from the video for playing video at different speeds, proportional to the likelihood of data to the query. It employs a similarity-based approach under generative models, that are based on the spatial layout of the video objects and trained on query clips. The generative models separate and balance different causes of variability of the objects in videos, like occlusions, appearance changes and motion.

Two distinct papers are concerned with video segmentation into semantically valuable video segments, and their use for creating video summaries. The first one, "*Motion-based selection of relevant video segment for video summarization*" by N. Peryard and P. Bouthemy proposes a two-stage approach for extracting interesting events based on low-level features. In the first stage, temporal segmentation of the video is achieved by detecting changes in the dominant image motion (i.e. camera motion) by means of an affine model. In the second stage, shots are classified using a statistical representation of the residual motion in the scene (the temporal co-occurrence of local motions) and a supervised classifier based on a Gibbsian model and a Maximum.-Likelihood estimator.

The second paper "*Extraction of film takes for cinematic analysis*" by B.T. Truong et al. provides video summarization after extraction of higher level features. It defines video takes as clusters of consecutive shots grouped at high level ("*one uninterrupted run of the camera to expose a series of frames*"). In this case, shot detection is not based on motion but rather on color analysis (by histogram differencing). Shots are then processed to filter out "action driven scenes". The remaining "drama driven scenes" are clustered into film takes based on shot similarity measures. Takes extraction is very useful both for selective video access and content summarization and for shot flow creation for the extraction of cinesthetic elements used in films, or for non-linear browsing of videos.

Finally, the paper, "*An integrated framework for semantic annotation and transcoding*" by M. Bertini et al. proposes automatic video annotation at the frame level. Meaningful highlights and objects of interests are extracted from not edited videos taken with a single main camera. Semantics extracted is classified into classes of relevance, that group together events and/or objects that have the same degree of importance for the user. In this paper, annotation is put in relationship with the process of transcoding. Different techniques of selective content-based transcoding of MPEG-2 videos are discussed. Different weights are associated to different classes of relevance and used to adjust the quantization factors so as

to save bandwidth in the case of non-relevant entities and improve the quality of display for the relevant ones.

Rita Cucchiara
Alberto Del Bimbo
Guest Editors