Using Knowledge Representation Languages for Video Annotation and Retrieval

M. Bertini, G. D'Amico, A. Del Bimbo, and C. Torniai

D.S.I. - Università di Firenze - Italy
{bertini, damico, delbimbo, torniai}@dsi.unifi.it

Abstract. Effective usage of multimedia digital libraries has to deal with the problem of building efficient content annotation and retrieval tools. In particular in video domain, different techniques for manual and automatic annotation and retrieval have been proposed. Despite the existence of well-defined and extensive standards for video content description, such as MPEG-7, these languages are not explicitly designed for automatic annotation and retrieval purpose. Usage of linguistic ontologies for video annotation and retrieval is a common practice to classify video elements by establishing relationships between video contents and linguistic terms that specify domain concepts at different abstraction levels. The main issue related to the use of description languages such as MPEG-7 or linguistic ontologies is due to the fact that linguistic terms are appropriate to distinguish event and object categories but they are inadequate when they must describe specific or complex patterns of events or video entities. In this paper we propose the usage of knowledge representation languages to define ontologies enriched with visual information that can be used effectively for video annotation and retrieval. Difference between content description languages and knowledge representation languages are shown, the advantages of using enriched ontologies both for the annotation and the retrieval process are presented in terms of enhanced user experience in browsing and querying video digital libraries.

1 Introduction and Previous Work

An ontology is a formal and explicit specification of a domain knowledge, typically represented using linguistic terms: it consists of concepts, concept properties, and relationships between concepts.

Several standard description languages for the expression of concepts and relationships in domain ontologies have been defined in the last years: Resource Description Framework Schema (RDFS), Web Ontology Language (OWL) and, for multimedia, the XML Schema in MPEG-7. Using these languages metadata can be fitted to specific domains and purposes, yet still remaining interoperable and capable of being processed by standard tools and search systems.

Ontologies can effectively be used to perform semantic annotation of multimedia content. For video annotation this can be done either manually, associating the terms of the ontology to the individual elements of the video, or automatically, by exploiting results and developments in pattern recognition and image/video analysis. In this latter case, the terms of the ontology are put in correspondence with appropriate knowledge models that encode the spatio-temporal combination of low and mid level features.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 634–646, 2006.

Once these models are checked, video entities are annotated with the concepts of the ontology; in this way, for example in the soccer video domain, it is possible to classify highlight events in different classes, like *shot on goal*, *counter attack*, *corner kick*, etc.

Examples of automatic semantic annotation systems have been presented recently, many of them in the application domain of sports video. Regarding the analysis of soccer videos we can cite [1] where MPEG motion vectors, playfield shape and players position have been used with Hidden Markov Models to detect soccer highlights. In [2] Finite State Machines have been employed to detect the principal soccer highlights, such as shot on goal, placed kick, forward launch and turnover, from a few visual cues. Yu et al. [3] have used the ball trajectory in order to detect the main actions like touching and passing and compute ball possession statistics for each team; a Kalman filter is used to check whether a detected trajectory can be recognized as a ball trajectory.

In all these systems model based event classification is not associated with any formal ontology-based representation of the domain. Domain specific linguistic ontology with multilingual lexicons, and possibility of cross document merging has instead been presented in [4]. In this paper, the annotation engine makes use of reasoning algorithms to automatically create a semantic annotation of soccer video sources. In [5], a hierarchy of ontologies has been defined for the representation of the results of video segmentation. Concepts are expressed in keywords and are mapped in an *object ontology*, a *shot ontology* and a *semantic ontology*.

The possibility of extending linguistic ontologies with multimedia ontologies, has been suggested in [6] to support video understanding. Differently from our contribution, the authors suggest to use *modal keywords*, i.e. keywords that represent perceptual concepts in several categories, such as visual, aural, etc. A method is presented to automatically classify keywords from speech recognition, queries or related text into these categories. Multimedia ontologies are constructed manually in [7]: text information available in videos and visual features are extracted and manually assigned to concepts, properties, or relationships in the ontology. In [8] new methods for extracting semantic knowledge from annotated images is presented. Perceptual knowledge is discovered grouping images into clusters based on their visual and text features and semantic knowledge is extracted by disambiguating the senses of words in annotations using WordNet and image clusters. In [9] a Visual Descriptors Ontology and a Multimedia Structure Ontology, based on MPEG-7 Visual Descriptors and MPEG-7 MDS respectively, are used together with domain ontology in order to support content annotation. Visual prototypes instances are manually linked to the domain ontology. An approach to semantic video object detection is presented in [10]. Semantic concepts for a given domain are defined in an RDF(S) ontology together with qualitative attributes (e.g. color homogeneity), low-level features (e.g. model components distribution), object spatial relations and multimedia processing methods (e.g. color clustering) and rules in F-logic are used for detection on video objects.

Despite of the difficulty of including pattern specifications into linguistic ontologies, classification at the pattern description level can be mandatory, in many real operating contexts. Events that share the same patterns can be represented by *visual concepts*, instead of linguistic concepts, that capture the essence of the event spatio-temporal development. In this case high level concepts expressed through linguistic terms, and pattern