MOM: Multimedia Ontology Manager. A Framework for Automatic Annotation and Semantic Retrieval of Video Sequences

Marco Bertini, Alberto Del Bimbo
Carlo Torniai
Università di Firenze - Italy
bertini,delbimbo,torniai@dsi.unifi.it

Rita Cucchiara, Costantino Grana
Università di Modena e Reggio Emilia - Italy
cucchiara.rita,grana.costantino@unimore.it

ABSTRACT

Effective usage of multimedia digital libraries has to deal with the problem of building efficient content annotation and retrieval tools. MOM (Multimedia Ontology Manager) is a complete system that allows the creation of multimedia ontologies, supports automatic annotation and creation of extended text (and audio) commentaries of video sequences, and permits complex queries by reasoning on the ontology.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.2.4 [**Systems**]: Multimedia databases

General Terms

Algorithms

Keywords

Video databases, Video annotation, Content-based retrieval, Multimedia ontology

1. INTRODUCTION

The basic idea behind multimedia ontologies is that the concepts and categories defined in a traditional linguistic ontology are not able to fully describe the diversity of visual events and elements that are in a video. In fact although linguistic terms are appropriate to distinguish event and object categories, they are inadequate when they must describe specific patterns of events or video entities. To this end high level concepts, expressed through linguistic terms, and pattern specifications, represented instead through visual or auditory concepts, both should be organized into new extended ontologies that couple linguistic terms with visual/audio information. The possibility of extending linguistic ontologies has been suggested in [1], [2], [3], [4].

MOM (Multimedia Ontology Manager) is a new system, which has been developed according to the principles and concepts of pictorially enriched ontologies, as defined in [4], [5], that supports dynamic creation and update of multimedia ontologies, provides facilities to automatically perform annotations and create extended text (and audio) commentaries of video sequences, and allows complex queries on video databases, based on the ontology itself. The MOM framework has been developed in the VAPEON project, as part of the activities of the DELOS Network of Excellence on Digital Libraries (Contract G038-507618) in the Information Society Technologies (IST) Program of the European Commission.

In the DEMO session we will show the complete MOM system under operation for the soccer video application domain, expounding: how ontologies that include visual concepts are created; how clusters of visual concepts are updated, as new examples are presented to the system; how new clips can be annotated with high level concepts of the ontology and text/audio commentaries are obtained automatically; how complex temporal queries can be performed over the database, using the temporal and semantic relationships defined in the ontology. In the following we expound some basic principles and technical details of the implementation. More will be explained and shown in the presentation.

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.

2. MULTIMEDIA ONTOLOGY CREATION

Multimedia ontologies as created from MOM, are expressed in the OWL standard. The linguistic part of the ontology is composed by a number of classes, that express the main concepts of the domain (f.e. actors, objects, facts and actions, highlights...) and their relationships. The extended multimedia ontology is created by linking video sequences as instances of concepts in the linguistic ontology, and performing an unsupervised Fuzzy C-means clustering of the instance clips. Visual features that are used for clustering are both generic visual attributes (e.g. trajectories, motion fields, edge and color histograms computed from image data...) and domain specific descriptors (e.g. spatio-temporal feature combinations...) that qualify special events. The centers of the clusters are regarded as visual concepts, each representing a specific pattern in which a fact or event can manifest. A special class *Undetected event/fact*) is also created, that holds all the clips that are not classified as concept instances up to a pre-defined confidence.

Fig. 1 shows an example of multimedia ontology for the soccer domain. The ontology includes concepts like *crowd cheering*, or *referee action*, such that their occurrences (e.g. crowd after a goal, crowd in a normal condition and crowd during attack action; referee yellow card, referee red card and referee fault signaling) can be easily distinguished each other, using generic features. Other more complex events like *highlights*, or *player actions* use instead more complex and domain-specific descriptors. The generic visual features used are: the color histogram (256 bins, in the HSV color space); the spatial color distribution (64 bins, obtained as the mean of colors in the YCbCr color space, for a 8x8 superimposed grid); the DCT coefficients (64 elements, according to the MPEG-7 specification for the Color Layout Descriptor); the average motion vectors (4 elements, obtained as the median of the MPEG motion vectors in each quarter of frame). The domain-specific descriptors used to model soccer highlights are: the playfield area; the number of players in the upper/lower part of the playfield; the camera motion intensity, direction and acceleration. For each clip, two feature vectors are created, respectively with six and four distinct components, one for each descriptor used, each component being a vector of as many elements as the number of frames in the clip, holding the values of the descriptor for each frame.

Composite concept patterns that are defined from temporal and

semantic relations between the concepts in the ontology, and that are characteristic of the application domain, can also be included in the ontology to be used to ease annotation of long video sequences and express complex queries.

3. AUTOMATIC VIDEO ANNOTATION

MOM allows effective automatic annotation of video clips with high level concepts. Annotation is performed at two distinct levels. At the clip level, the video sequence is segmented into clips, and each of them is annotated by checking its similarity with the visual concepts of the ontology. As the similarity with a particular visual concept is assessed, then higher level concepts linked to it in the ontology are immediately associated with the clip. Newly annotated video clips are associated with the existing clusters. As a result, the centers of the clusters can change, and clips in the *Undetected event/fact* cluster must be re-analyzed to check if some of them can be associated to the new clusters. It is worth to notice that due to this mechanism the ontology has a static linguistic part (concepts and their relations are fixed and reflect the agreed description of the domain) and a visual part which is instead subjected to changes, (the centers of clusters - the visual concepts - change as new knowledge is presented to the system).

At the sequence level, composite concept patterns and the RACER description logic reasoner can be used to annotate a sequence of clips with some pre-defined articulated sentence. In fact, MOM allows to check if a video sequence contains a sequence of clips such that verifies one of the composite concept patterns pre-defined. Just as an example, let's consider a soccer video sequence containing the ordered succession of highlights and events: placed kick, forward launch, shot on goal and score change. Once each of the clips is separately annotated and their temporal order is assessed, the whole sequence can be annotated as an instance of the concept pattern *Video with Scored Goal*.

By reasoning on composite concept patterns, and using also the clip visual descriptors (which provide information about the playfield zone, the motion intensity of the action and the number of players involved), more extended commentaries can be created and associated automatically to video sequences (see Fig. 2). The commentary stored in srt format can be played as a video subtitle or made available as a simple text/audio file to be accessed through the web or downloaded to a mobile device.

4. SEMANTIC QUERYING

Video clips that have the same visual patterns can be easily retrieved by checking the similarity between the query example and the clip instances of the ontology. Moreover, by reasoning over the ontology, MOM allows the expression and solution of complex queries. Queries are expressed in nRQL. As an example, queries

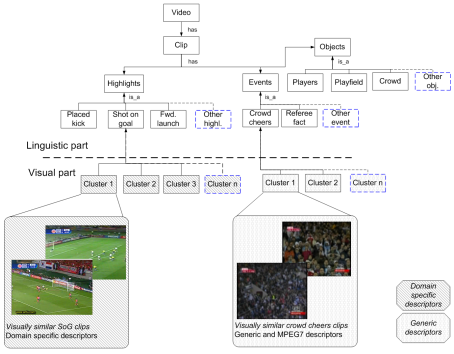

Figure 1: simplified view of soccer domain ontology with clusters of visual instances

Figure 2: Automatically generated clip subtitle

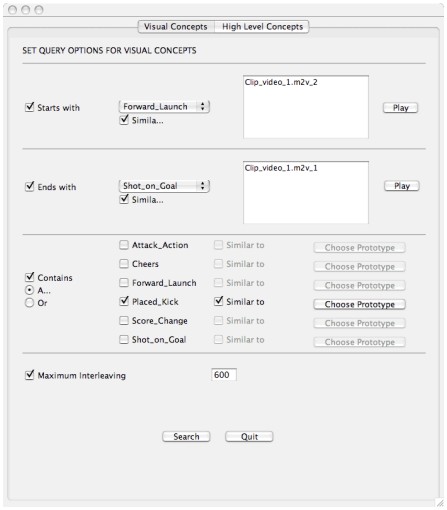

Figure 3: MOM query interface and query expression: search for sequences starting with a forward launch and finishing with a shot on goal, and including a placed kick, where all highlights must be similar to selected examples and the maximum delay between the highlight clips is 600 seconds.

such as: "*find a video sequence with a scored goal after a fault shortly followed (less than 10") by a fast attack action*", can be easily solved by the MOM reasoner. For each instance of the *Clip* and *Video* class, the reasoner evaluates the inferred types, and associates to each instance the proper types considering the actions detected or the patterns contained and finds the video sequences such that their clips verify the special pattern of the query. Visual concepts can also be used as examples in these queries to have more precise indication of the query target. With MOM we can retrieve video sequences with some specific highlight pattern and highlights similar to selected visual concepts. An example of such a query is shown in Fig. 3.

5. REFERENCES

[1] A. Jaimes and J. Smith. Semi-automatic, data-driven construction of multimedia ontologies. In *Proc. of IEEE ICME*, 2003.

[2] J. Strintzis, S. Bloehdorn, S. Handschuh, S. Staab, N. Simou, V. Tzouvaras, K. Petridis, I. Kompatsiaris, and Y. Avrithis. Knowledge representation for semantic multimedia content analysis and reasoning. In *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, Nov. 2004.

[3] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, and M. G. Strintzis. Knowledge-assisted semantic video object detection. *IEEE TCSVT*, Vol. 15, No. 10, Oct. 2005.

[4] M. Bertini, R. Cucchiara, A. Del Bimbo and C. Torniai, Video Annotation with Pictorially Enriched Ontologies. In *Proc. of IEEE ICME* , 2005.

[5] M. Bertini, A. Del Bimbo and C. Torniai, Enhanced Ontologies for Video Annotation and Retrieval. In *Proc. of ACM MIR*, 2005.