Semantic annotation and retrieval of video events using multimedia ontologies

Andrew D. Bagdanov, Marco Bertini, Alberto Del Bimbo, Giuseppe Serra, Carlo Torniai
Università di Firenze - Italy
Via S. Marta, 3 - 50139 Firenze
{bagdanov,bertini,delbimbo,serra,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 this paper Multimedia Ontologies, that include both linguistic and dynamic visual ontologies, are presented and their implementation for soccer video domain is shown. The structure of the proposed ontology itself, together with reasoning, can be used to perform higher-level annotation of the clips, to generate complex queries that comprise actions and their temporal evolutions and relations and to create extended text commentaries of video sequences.

1. Introduction

The importance of non-textual media in the context of digital libraries has grown steadily in recent years. Digital libraries today are expected to include, and provide effective access to, a broad range of heterogeneous media including text, video, audio and graphics. Digital video is the media that is of greatest relevance due to the inability of digital librarians to cope with the sheer quantity produced daily by broadcasters, media companies, government institutions and individuals for news, personal entertainment, educational, and institutional purposes. Video poses significant challenges to digital librarians due to the size of files, the temporal nature of the medium, and the lack of bibliographic methods that leverage non-textual features.

Put simply, digital video does not fit into the traditional paradigms used by librarians to collect, categorize and index information. Librarians have indexed video collections with textual metadata: the producer name, the date and time, and a few linguistic concepts that summarize video content at semantic level. In this way video can be accessed almost in the same way as textual documents. Effective examples of retrieval by content of video clips using textual keywords have been presented for news [6, 10, 14] and sports video domains [7, 19]. This brand of manual annotation, however, is labor-intensive, expensive, prone to error and imprecision, and fails to capture the dynamic aspects of the medium.

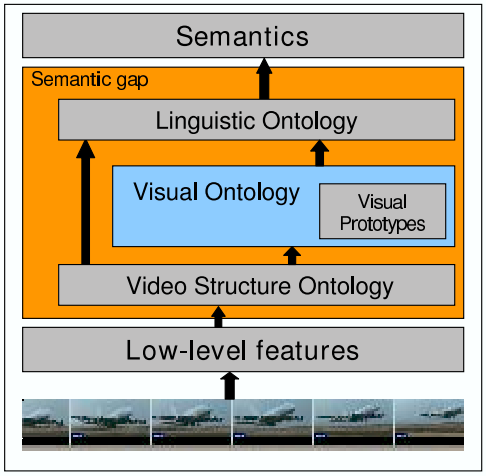


Figure 1. Several levels of ontologies are used to bridge the semantic gap between data and semantics.

Richer annotation of digital video requires that more complex linguistic structures be used to represent knowledge about video at a deeper semantic level. Ontologies are defined as the representation of the semantics of terms and their relationships. They consist of concepts, concept properties, and relationships between concepts, all expressed in linguistic terms [8]. For the digital video domain, ontologies can describe either the video content domain, in which case they are static descriptions of entities and highlights present in the video and their relationships, as codified by human experience. Or, they can describe the structure of the media, i.e. the component elements of the video, the operations allowed on its parts, and the low-level video descriptors that characterize their content.

Traditional domain ontologies, whether in the form of textual metadata or linguistic abstractions and relations de-

0-7695-2997-6/07 $25.00 © 2007 IEEE
DOI 10.1109/ICSC.2007.30

fined on primitive video elements, are substantially inadequate to support complete annotation and retrieval by content of video documents. A single, perceivable concept in a domain ontology can manifest itself in so many different instantiations of binary data (in the form of a digital video) that enumeration of all instances is effectively impossible. The level of detail at the resolution of binary data makes precise description of a visual concept in linguistic terms impractical when not impossible. The primary reason for this is the vast divide separating such linguistic abstractions from the low-level, binary representation of the digital video medium. This divide is commonly referred to as the *semantic gap* and is depicted in figure 1 along with the ontological components we use to bridge it.

A simple example helps to illustrate this point. Consider the problem of detecting *attack action* highlights in soccer videos. Among the various types of highlights possible in soccer videos, those that can be classified as *attack actions* can manifest themselves according to many different patterns of high-level actions, features, and events. These patterns may differ in the playfield zone where the action takes place, the number of players involved, the player motion direction, the speed and acceleration of the key player, etc. Although we are able to distinguish between them, and mentally categorize the patterns of attack actions into distinct classes, to express each pattern in linguistic terms would require a complex sentence, explaining the way in which the action develops. Such a sentence indeed should express the translation of our visual memory of the action into a conceptual representation where concepts are related in complex ways according to spatio-temporal constraints. In this translation we typically make a synthesis that retains only the presumed most significant elements and much precision in visual detail is lost, some facts will are omitted, and, probably most importantly, the most appropriate words to distinguish one pattern from another are not used.

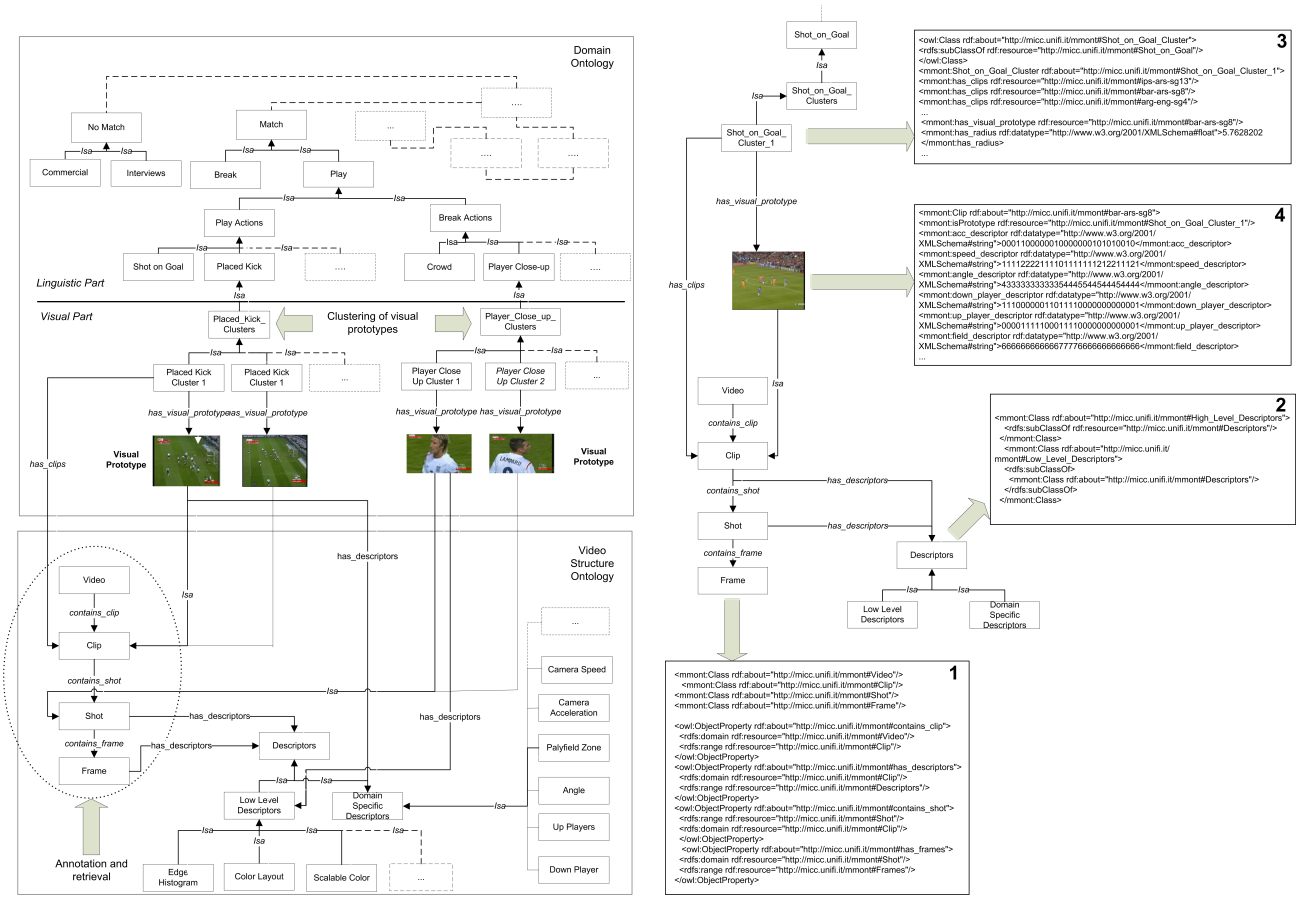
Even at the relatively high level of description used in this example, i.e. in terms of players, field position, and dynamics, the problem of characterizing visual concepts linguistically is extremely daunting. To begin bridging the semantic gap, ontologies can be enriched to include structural video information and visual data descriptors, growing the representation upwards, in a sense. In [17], a Visual Descriptors Ontology and a Multimedia Structure Ontology, respectively based on MPEG-7 Visual Descriptors and MPEG-7 Multimedia Description Schema, are used together with a domain ontology in order to support video content annotation. In [13], a hierarchy of ontologies was defined for the representation of the results of video segmentation. Concepts were expressed in keywords using an *object ontology*: MPEG-7 low-level descriptors were mapped to intermediate level descriptors that identify spatio-temporal objects.

To resolve the inadequacies of traditional linguistic ontologies in describing these complex phenomena, the need for both *conceptual* and *perceptual* abstractions must be recognized. Jaimes et al. [11] suggested that concepts that relate to perceptual facts be categorized into classes using *modal keywords*, i.e. keywords that represent perceptual concepts in several categories. This is a key observation that can be used to great advantage once we have a method to classify phenomena into perceptual categories of the domain.

In all these solutions, structural and media information are still represented through linguistic terms and fed manually to the ontology. In [15] three separate ontologies modeling the application domain, the visual data and the abstract concepts were used for the interpretation of video scenes. Automatically segmented image regions were modeled through low-level visual descriptors and associated with semantic concepts using manually labeled regions as a training set. In [5], qualitative attributes that refer to perceptual properties like color homogeneity, low-level perceptual features like model components distribution, and spatial relations were included in the ontology. Semantic concepts of video objects were derived from color clustering and reasoning. In [16] the authors have presented video annotation and retrieval based on high-level concepts derived from machine learned concept detectors that exploit low level visual features. The ontology includes both semantic descriptions and structure of concepts and their lexical relationships, obtained from WordNet.

In this paper we present a solution for the definition, implementation, and application of multimedia ontologies for the soccer video domain. The multimedia ontology is defined using the Web Ontology Language (OWL). It includes concepts at the abstract, linguistic level as well as at the perceptual level. Most importantly, it contains concepts that establish relationships between linguistically defined facts and their perceptual manifestations in digital video. In this way, the logical structure of the high-level domain ontology, defined in terms of linguistic concepts and relations, can be used to induce structure in the unorganized, binary data at the other end of the semantic gap.

The organization of the paper is as follows. In the next section the structure and the definition of a multimedia ontology for the soccer video domain is presented. In section 3 the descriptors used to identify and model the multimedia extensions of the domain specific concepts of the ontology are briefly presented. In section 4 the usage of the multimedia ontology for automatic video annotation and creation of extended video commentaries are presented. We conclude in section 5 with a discussion of the main results and observations of the application of multimedia ontologies to the understanding of digital video.



(a) (b)

Figure 2. Multimedia ontology for soccer videos: (a) partial view of the principal classes, relations and instances; (b) OWL definition of the principal parts of our multimedia ontology.

2. A multimedia ontology for soccer videos

In multimedia ontologies the linguistic and perceptual parts have substantially different characteristics. The linguistic part of the ontology embeds permanent, self-contained, objective items of exact knowledge, that can be taken as concrete and accepted standards by which reality can be considered. The perceptual part, on the other hand, includes concepts that are not abstractions, but are mere duplicates of things observed in reality. Figure 2(a) shows the important parts of our multimedia ontology for soccer video. It is composed of two main parts: the soccer domain ontology and the video structure ontology.

The domain ontology contains all the concepts and relations defining the soccer domain. It includes the high level concepts (expressed in linguistic terms) that name the entities and facts of soccer, and the perceptual elements that model the visual patterns of the specializations of the concepts of the linguistic part. The linguistic part of the domain ontology contains concepts that correspond to observable, perceptual facts and events that exist or occur in the real world. Concepts modeled in our soccer domain ontology include `Shot_on_Goal` [1], `Placed_Kick`, `Attack_Action`, and `Forward_Launch`. Concepts can also be thought of "highlight" events in soccer video, i.e. events that are of interest in and of themselves, or play an important role in compound events.

Each linguistic concept in the domain ontology is associated with a corresponding *visual concept*. To account for the many perceptually different patterns in which patterns manifest themselves, visual concepts are clustered according to the similarity of their spatio-temporal patterns. For each cluster, a visual prototype is obtained as the representative element for that class. The visual prototype acts as a bridge between the domain ontology and the video structure ontology. A visual prototype is a structural element of the video (either a clip, a shot, a frame or part of a frame) and is linked to a linguistic concept in the domain ontology (a play action, a player). This link plays an important role for au-

[1] Concepts from the ontology are typeset using the `typewriter` font.

tomatic annotation of new video sequences and for retrieval by content as explained in section 4.

The video structure ontology describes the component elements of a video: clips, shots, frames, etc. For frames and shots it defines the descriptors of the visual content at low-intermediate level. It should be noted that a large portion of this fragment of the ontology is generic and generalizes naturally to other domains. They are not expressed in linguistic terms but rather by a set of values that are attributed to a few features that are agreed to model the perceptual content of the concept. These values have substantially different dynamics than linguistic terms and in a sense establish the basic elements of discourse for describing knowledge at the semantic level in the domain.

The OWL definitions of the principal elements that support bridging between structure and domain ontologies are shown in figure 2(b). Box 1 includes the definitions of the key elements that compose the structure of a video document (`Video`, `Clip`, `Shot` and `Frame` classes). From top to bottom they are related through the *contains* relation. The definitions of the `Descriptor` class and its subclasses are shown in box 2. They respectively model low level descriptors and domain-specific visual descriptors of `Shot` and `Frame` video components Box 3 contains the definition of clusters of perceptual facts. A `Shot_on_Goal_Cluster` is shown for the sake of clarity. Each cluster class, which represents cluster of perceptual facts in the ontology, has the following properties defined on it:

property	description
`has_clips`	An enumeration of all `Clips` in the cluster
`has_visual_prototype`	The cluster center, an instance of `Clip` in this case
`has_radius`	size of the visual concept cluster

Domain specific visual descriptors and their particular values for the `Shot_on_Goal_Cluster_1` visual prototype are listed in box 4. The `isPrototype` property of a structural element, e.g. a `Clip`, and the `has_visual_prototype` property on perceptual concepts establish the link between high-level domain ontology concepts and structural elements of the video document.

3. Modeling perceptual concepts

Descriptors of perceptual concepts regard image regions in frames (for entities and subjects) or sequences of frames (for scenes, highlights and events). They include color, texture or pattern descriptors, and their temporal distribution. We model perceptual observations in soccer videos at three levels: the scene, highlights that can occur in a scene, and scene subjects (i.e. the actors in the scene).

3.1. Scene modeling

We model four of the most common types of scenes in soccer videos:

- **play:** which is visually characterized by a few large, homogeneous color regions and long playfield lines;
- **player closeup:** in which a face appears distinctly while the background is of homogeneous color and/or blurred;
- **player medium view:** where several player faces are distinguishable and their bodies form relatively large regions of uniform color; and
- **crowd:** in which individuals are not clearly evidenced, but rather appear as a texture.

These observations suggest that global color features, layout of homogeneous colour areas, edge and shape features can be used differentiate these scene types. We exploit the following generic attributes defined in the MPEG-7 standard for multimedia content description, applied to each frame: scalable color descriptor, color layout descriptor and edge histogram descriptor.

3.2. Soccer highlight modeling

Soccer highlights are detected only for play scenes. They are distinguished on the basis of the spatio-temporal combination of a reduced set of visual features: the camera motion direction and intensity (approximately modeling the key players' motion); the playfield zone; the number of players in the upper and lower part of the playfield. The features are obtained from the compressed and un-compressedvideo domain, as described in [3].

3.3. Scene subject modeling

Players, referees, and trainers are important subjects for semantic annotation and content-based retrieval of video clips and episodes. Generally speaking, however, identification of these subjects is a hard task: occlusions, fast motion and low resolution prevent reliable face detection and recognition. Though in practice only close-up views are useful, problems persist due to the large variation in face pose and expression.

Nevertheless, detection of these subjects is an important indicator of the types of events we are interesting in modeling. In close-up scenes we observe that the faces of individuals that perform important actions are typically framed with part of the body included, therefore showing the team

jersey and printed data such as the player number and name. Such close-ups additionally often have superimposed text captions with the player name. This suggests that players, referee and trainer can be automatically identified by exploiting face information together with the information extracted from the jersey color, number and text, and/or the superimposed text captions.

Face sequence detection To detect faces, we use a slightly modified version of the AdaBoost face detector of [18]. Face recognition is performed using local features to describe a face, and sets of poses to describe a person [1].

Text detection and recognition Two different types of text usually appear in soccer videos: superimposed text, that usually contains information about the teams and player names, and the player jersey numbers and names.

The style of the superimposed captions is different for every broadcaster, and has changed often throughout the years. Official rules of most important soccer organization (like UEFA and FIFA) state that the front of player jerseys are decorated with numbers of a specific size. Player numbers are always in the range from 1 to 22, and each number is assigned to a player for the entire duration of the tournament.

Detection of superimposed text does not require any knowledge or training on superimposed captions or scene text features, and does not use temporal redundancy for the text detection and extraction. Image corners and MSER points are used for this task [2].

4. Automatic video annotation and retrieval

In this section we describe some of the applications that are enabled by modeling the soccer video domain at the conceptual level with relationships establishing links to their lower-level, perceptually representative manifestations.

4.1. Annotation of highlights and events

The multimedia ontology can be effectively used to perform automatic video annotation with high level concepts that describe what is occurring in the video clips [3]. This is made by checking the similarity of video content with the visual prototypes included in the ontology. If similarity is assessed with a particular visual concept then also higher level concepts in the ontology hierarchy, related to that visual concept, are associated to the video.

4.2. Reasoning engine

Use of OWL DL for the multimedia ontology description allows us to use inference to extend and exploit the knowledge in the ontology. We use RacerPro [9] description logic (DL) reasoner to infer new knowledge that is used to refine annotation, which is necessary to model the type of high-level event descriptions we are interested in detecting automatically. The Jena APIs [12] have been used to pose queries with SPARQL [4].

4.3. Annotation of episodes with composite events and high level concepts

Some high level concepts that of interest can not be detected and recognized from the visual descriptors or detection of the entities in a clip. They are determined, rather, by a temporal sequence of visual descriptors. For example in soccer videos the detection of a scored goal can not be recognized by the analysis of the visual descriptors alone. In fact, note even ball tracking and recognition of it crossing the goal line are sufficient, since a goal is assessed only after the referee decision. In other cases certain sequences can be related to a semantic meaning depending on the intent of the video producer; an example is the case of soccer fouls, in which the action is followed by a set of clips that contain player close-ups and medium views that show the injured and the offending players, and their teammates. Finally there can be concepts that are combinations of simpler concepts, such as a `Counter_Attack` action that is obtained as a sequence of two attack actions developed by different teams.

These cases can be regarded as patterns of events, and can be discovered through inference, where a linear sequence of visual concepts is inferred through the `followedBy` relation in the To this end we have identified some meaningful *patterns* of actions and events. For example, attack actions that terminate with a scored goal are always followed by crowd cheering in crowd scenes, player close ups and medium views, and superimposed text reporting a score change. A pattern for `Scored_Goal` can be formally defined in the ontology as subclass `Video_with_Scored_Goal`, that contains this combination of facts:

- `Forward_Launch` followed by `Shot_on_Goal` followed by (`Crowd` or `Player_Closeup` or `Medium_View`) followed by `Score_Change`; or
- `Placed_Kick` followed by `Player_Closeup` followed by `Score_Change`; or
- `Shot_on_Goal` followed by (`Crowd` or `Player_Closeup` or `Medium_View`) followed by `Score_Change`.

In this way temporal constraints are used to distinguish meaningful sequences of events, so as to avoid considering

actions that are not related to each other. A video sequence is classified as `Video_with_Scored_Goal` by the reasoner if the ordered sequence of clips contained in it corresponds to the pattern definition described above.

4.4. Automatic creation of extended commentaries

Patterns and perceptual descriptors of clips (like the playfield zone, the motion intensity of the action and the number of players, the player names, etc.) allow the automatic construction of extended commentaries of video sequences. To obtain automatic commentaries of a video stream, a set of basic sentence templates are associated with the high-level event types that can be asserted or inferred for a video clip. These sentences are picked up randomly by a commentary generator engine to give a variety of expressions and make the commentary more realistic. More precise descriptions are obtained by considering the values of the visual descriptors of the clips that have been assigned the type corresponding to a sentence template. Considering for example motion intensity, or the number of players and the playfield zone descriptors, the sentences are refined by means of adverbs and adjectives determined by visual descriptors such as: the playfield area in which the action is taking place, the amount of player in the playfield area, and the speed of the action. *A priori* knowledge and other asserted metadata is also used, such as: the name of the home team, the name of the visitor team, the current score of the match, and the current time of the match.

Sentence templates are represented in a simple XML format with certain tags representing terms to be inferred from specific instances of a type of fact in the ontology, or a randomly selected modifier of a type. For instance a simple sentence for attack action clips is represented by:

```
<Sentence type="Attack_Action">
<Home_Team/>
<Attack_Action/>
<Modifier_Speed/>
<Modifier_Players/>
<Modifier_Playfield_Zone/>
</Sentence>
```

Tags values are set taking into account the values of visual descriptors and the a priori knowledge stored in the ontology so that the following commentary can be obtained:

> *England surprises the defense with a very fast shot from a crowded midfield toward the goal area.*

If a scored goal pattern is recognized within the clip the values of tags will be different resulting as in the following commentary

> *England scores an incredible goal with a fast shot from the crowded midfield area.*

Commentaries are stored in SRT format and presented as text subtitles (see Fig. 3) or stored as a text file that can be accessed through the web or downloaded to a mobile device.

4.5. Query and retrieval of video clips or episodes

Exploitation of the multimedia ontology permits queries that combine visual concepts (i.e. special patterns of highlight visual features) with high level concepts (such as team names, players, date of the match, locations, etc) and temporal constraints to retrieve even long video sequences obtained from the concatenation of shots.

Queries can be expressed by means of a simplified mask that allows to define combinations of video highlights and to browse linguistic ontology for the selection of concepts of interest. The complexity of the query expression is hidden to the final user and the SPARQL syntax is dynamically generated by the graphical user interface.

Reasoning is used in the query process in order to compute inferred types for clips and shots and find the patterns occurrences in videos. In particular, for the inferred type computation, clips or shots are classified as the proper type ("Clip with Attack Action", "Clip with Shot on Goal", "Shot with Player Close Up"...) depending on the kind of highlights they contain. Similarly, inference on video instances classifies every sequence according to the pattern it may contain: for instance a video will be classified by the reasoner as "Video with Scored Goal" if it contains one of the required sequence of clip types defined in the Scored Goal pattern.

In Fig. 4 an example of query requesting shot on goal with play patterns similar to a sample shot on goal clip (row 1 of Fig. 5) is shown. Results are shown in Fig. 5. For each clip only a small number of frames are shown. In this case the reasoner is required to compute only the inferred type "Clip with Shot on Goal".

A more complex query that can be performed is to request a temporal sequence of events that represent composite events. An example is a query that requests a sequence that contains an "interesting" shot on goal defined as an attack action, followed shortly by a shot on goal and a clip that contains crowd or players' closeups. The inference engine computes the inferred types ("Clip with Shot on goal", "Clip with Attack Action" and "Shot with Crowd"), checks their occurrences in the same video according to the defined order and temporal constraints, and presents the results.

If a pattern for typical complex events has been defined in the ontology, such as the scored goal or a foul pattern described in Sect.4.3, it is possible to query directly for it.

Figure 3. Automatically generated subtitle for a sequence of clips.

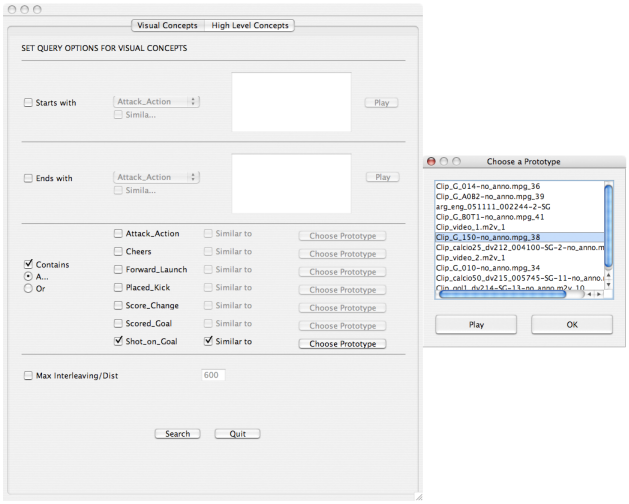


Figure 4. Query example: search for a shot on goal similar to a visual prototype.

Finally it is possible to combine highlights sequences, visual similarity and high level concepts in the same query. For example, it is possible to search for sequences starting with a shot on goal similar to a visual prototype followed by placed kicks within 60 seconds. Results can be further constrained by requesting only events involving specific teams.

5. Discussion

In this paper we have described our ongoing work on multimedia ontologies that model knowledge about the soccer video domain at the linguistic, perceptual, visual, and media levels. Central to our description and implementation of multimedia ontologies are the relationships established between visual and linguistic concepts through visual prototyping, a process of clustering perceptual manifestations of visual concepts through which a *visual prototype* is designated as a representative of a particular class of visual phenomenon. These prototypes, which are instances of low level concepts and concept descriptors defined in a video structure ontology, establish the link between the binary, digital video domain and the logical, linguistic concepts defined in the high-level domain ontology.

We also illustrated the power of this type of representation though several example applications we have built. Automatic, semantic-level annotation of new video elements such as clips and entire videos is made possible though comparison with existing visual prototypes in the ontology.

Figure 5. Results of a query for shot on goal similar to the visual prototype of the first row.

Inferred properties on video elements allow for richer structure to be inferred directly in the video domain through inference of higher-level composite events defined as temporal sequences of primitive concepts. Finally, queries can be posed at the level of semantic descriptors, hiding the complexities of the low-level representation and description of digital video from the user.

Our multimedia ontology framework is currently being generalized to other application domains, including broadcast news and video surveillance video. New domains obviously require new linguistic abstractions, but we are investigating how to generalize abstractions at the video structure and visual domain level so that much of the existing ontology structure can be reused. This is especially desirable because most of the inference in our applications occurs at this level, and the clustering of visual elements into designated visual prototypes is driven by the these relational structure as well.

Temporal specification of high-level, composite events is also of critical importance. We have barely scratched the surface of what is possible with temporal specification and reasoning. Indeed, many of the interesting events that occur in most video domains cannot be described in terms of individual, atomic events that can be characterized by simple clustering of visual descriptors. Currently we use simple, linear sequences of events, but we are extending the temporal relationships in our ontology using the Allen interval algebra. This extension will enable much richer description and inference of high-level events.

Acknowledgments

This work is partially supported by the Information Society Technologies (IST) Program of the European Commission as part of the DELOS Network of Excellence on Digital Libraries (Contract G038-507618).

References

[1] L. Ballan, M. Bertini, A. Del Bimbo, and W. Nunziati. Automatic detection and recognition of players in soccer videos. In *In Proc. of the International Conference on Visual Information Systems*, June 2007.

[2] M. Bertini, A. Del Bimbo, and W. Nunziati. Automatic detection of players identity in soccer videos using faces and text cues. In *In Proc. of ACM Multimedia*, October 2006.

[3] M. Bertini, A. Del Bimbo, and C. Torniai. Automatic video annotation using ontologies extended with visual information. In *Proceedings of ACM Multimedia*, November 2005.

[4] W. W. W. Consortium. Sparql protocol and rdf query language. Technical report, W3C, http://www.w3.org/TR/rdf-sparql-query/, accessed 4 October 2006.

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

[6] A. Eickeler and S. Muller. Content-based video indexing of tv broadcast news using hidden markov models. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2997–3000, March 1999.

[7] A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7):796–807, July 2003.

[8] T. Gruber. Principles for the design of ontologies used for knowledge sharing. *Int. Journal of Human-Computer Studies*, 43(5-6):907–928, 1995.

[9] V. Haarslev and R. Möller. Description of the racer system and its applications. In *Proceedings International Workshop on Description Logics (DL-2001), Stanford, USA, 1.-3. August*, pages 131–141, 2001.

[10] A. Hauptmann and M. Witbrock. Informedia: News–on–demand multimedia information acquisition and retrieval. *Intelligent Multimedia Information Retrieval*, pages 213–239, 1997.

[11] A. Jaimes, B. Tseng, and J. Smith. Modal keywords, ontologies, and reasoning for video understanding. In *Int'l Conference on Image and Video Retrieval (CIVR)*, July 2003.

[12] B. McBride. Jena: Implementing the rdf model and syntax specification. In *In Proc. of the Second International Workshop on the Semantic Web - SemWeb'2001*, May 2001.

[13] V. Mezaris, I. Kompatsiaris, N. Boulgouris, and M. Strintzis. 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, 2004.

[14] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith. Video ocr for digital news archive. *IEEE International Workshop on Content–Based Access of Image and Video Databases CAIVD' 98*, pages 52–60, 1998.

[15] N. Simou, C. Saathoff, S. Dasiopoulou, E. Spyrou, N. Voisine, V. Tzouvaras, I. Kompatsiaris, Y. Avrithis, and S. Staab. An ontology infrastructure for multimedia reasoning. In *Proc. International Workshop VLBV 2005*, Sardinia (Italy), September 2005.

[16] C. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, (Pending minor revision), 2007.

[17] 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.

[18] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[19] X.Yu, C. Xu, H. Leung, Q. Tian, Q. Tang, and K. W. Wan. Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In *ACM Multimedia 2003*, volume 3, pages 11–20, Berkeley, CA (USA), 4-6 Nov. 2003.