# 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

specifications represented instead through visual concepts, can be both organized into new extended ontologies. In the following we will refer to them as *pictorially enriched ontologies*.

The basic idea behind pictorially enriched ontologies is that the concepts and categories defined in a traditional ontology are not rich enough to fully describe the diversity of all the visual events, and of their patterns, that normally are grouped in a same class and cannot support video annotation up to the level of detail of pattern specification. To a broader extent the idea of pictorially enriched ontologies can be extended to *multimedia enriched ontologies* where concepts that cannot be expressed in linguistic terms are represented by prototypes of different media like video, audio, etc.

This paper presents pictorially enriched ontologies, discusses a solution for their implementation for the soccer video domain and proposes a method to perform automatic soccer video annotation using these extended ontologies. The PE ontology creation process assigns multimedia objects to concepts and integrates the semantics described by the linguistic terms, while reasoning on the ontology adds a higher level of semantic annotation using the concepts relations, and allows to perform complex queries based on visual concepts and patterns of actions. The advantage of pictorially enriched ontologies is twofold:

- visual concepts allow to associate automatically occurrences of events or entities to higher level concepts by checking their proximity to visual concepts that are hierarchically linked to higher level semantics
- the unification in the same ontology both of specific domain concepts and their multimedia low and mid level descriptions allows the development of more user friendly interfaces for contend-based browsing and retrieval, using visual concepts and concepts relations.

The paper is organized as follows: an analysis of different ontologies standard languages is provided in Sect. 2. Creation of a pictorially enriched ontology for the representation of highlight patterns of soccer videos and the visual features extraction process are discussed in Sect. 3. Two algorithms that use the enriched ontology to perform automatic annotation are briefly presented in Sect. 4. In Sect. 5 is shown how the proposed ontology structure and ontology-based reasoning add a more refined annotation to the videos, allowing the retrieval of video content by mean of complex queries on the ontology. In Sect. 6 we discuss the preliminary results of the proposed system applied to soccer videos annotation. Finally, in Sect. 7 we provide conclusions and some future works.

## 2 Ontologies Standards

There are many differences between description languages based on XML (XML-Schema and MPEG-7) and knowledge representation languages (RDFS and OWL): both language categories can represent ontologies but with different capabilities of expressiveness and functionalities. MPEG-7 is a standard that has been built to define entities and their properties wrt the specific domain of multimedia content while RDFS

and OWL are languages that can define an ontology in terms of concepts and their relationships regardless of the domain of interest. The advantages of using MPEG-7 in multimedia domain is due to the fact that it has been designed to fully describe multimedia document structure, but at the same time it reflects the "structural" lack of semantic expressiveness of XML. In fact XML can state only few relations and most of them are not explicitly semantic since XML Schema can only define syntactic structure for a specific document: typically in an XML document only taxonomy relations are stated and most of them imply other high level semantic relation that are not directly expressed.

Knowledge representation languages extend the capability and the expressiveness of XML. RDFS can define an ontology in terms of concepts, properties and relationships of concepts without any restriction. OWL adds to RDFS the capability to refine concept definition and class restrictions. Both of them are flexible and extensible because they are not standard for a specific domain but they have been designed as general-purpose languages for domain independent knowledge description. Moreover knowledge representation languages can support usage of inference engines that can enrich the knowledge of a domain with the inferred knowledge. Due to these intrinsic characteristics MPEG-7 and RDFS/OWL can have different scope of utilization in multimedia domain. MPEG-7, as a standard provides all the necessary definition for structural description of multimedia content in particular for low-level descriptors such as color and edge histograms, texture descriptors, motion parameters, etc., but on the other hand it cannot be extended and can hardly include, for instance, high semantic structured description of video content.

Low-level descriptors can be very useful for description purpose, similarity assessment and annotation in simple domains but for effective semantic annotation of complex video content both low and mid level features have to be taken into account. It is possible to create new MPEG-7 Description Schemas, and also Descriptors, using the MPEG-7 Description Definition Language (DDL), but it has to be noted that this language, based on a variation of XML Schema, is not a modeling language, but is rather to be used to express the results of modeling. RDFS and OWL can easily be used to represent any kind of domain so they are suitable for describing both the structure of a multimedia content as well as its content in a structured semantic way. It is possible to include in a single RDF or OWL document a domain ontology referring to, for instance, the soccer domain and a multimedia ontology describing the structure of video in terms of segments, regions, etc.

There are several advantages in using ontologies specific standards as RDF and OWL instead of MPEG-7:

- it is easier to add mid-level audio-visual descriptors that are related to the domain that is being annotated.
- it is possible to express easily complex semantic entities and their relations;
- it is possible to use reasoning tools to add high-level annotation or perform queries related to semantic content, rather than low level audio-visual descriptors;

It has to be noted that the possibility to translate MPEG-7 into an ontology language such as RDF and OWL has been exploited to overcome the lack of formal semantics of the MPEG-7 standard that could extend the traditional text descriptions into machine

understandable ones. The first attempt that aimed to bridge the gap between the MPEG-7 standard and the ontology standards has been presented in [11] and [12]. In these works the first translation of the MPEG-7 MDS into RDFS has been shown. The resulting ontology has been also converted into DAML+OIL, and is now available in OWL. The ontology, expressed using OWL Full, covers the upper part of the Multimedia Description Schema (MDS) part of the MPEG-7 standard. It consists of about 60 classes and 40 properties. A methodology and a software implementation for the interoperability of MPEG-7 MDS and OWL has been presented in [13] and [14], developing from the previous work, and using OWL DL.

Another MPEG-7 Ontology is the one provided by the DMAG group at the Pompeu Fabra University[1]. This MPEG-7 ontology has been produced fully automatically from the MPEG-7 standard in order to give it a formal semantics. It is an OWL Full ontology, and aims to cover the whole standard and is thus the most complete one. It contains 2372 classes and 975 properties.

## 3 Pictorially Enriched Ontologies

As an example of pictorially enriched ontology we refer for the sake of clarity to Fig. 1, in which the linguistic and visual parts of the ontology are shown. The linguistic part is composed by the video and clip classes, the actions class and its highlights subclasses and an object class with its related subclasses describing different objects within the clips. In this example only placed kick, shot on goal and forward launch are shown.

The visual part is created adding to the linguistic part of the ontology the visual concepts as specializations of the linguistic concepts that describe the highlights. Visual concepts in the visual part are *abstractions* of video elements and can be of different types:

- *sequence* (the clip at the center of the cluster);
- *keyframes* (the key frame of the clip at the center of the cluster);
- *regions* (parts of the keyframe e.g. representing players);
- *visual features* (e.g. trajectories, motion fields, computed from image data ...).

Pictorially enriched ontologies are expressed using the OWL standard so that they can be shared and used in a search engine to perform content based retrieval from video databases or to provide video summaries.

The creation process of the pictorially enriched ontology is performed by selecting a representative set of sequences containing highlights described in the linguistic ontology, extracting the visual features and performing an unsupervised clustering. The clustering process, based on visual features, generates clusters of sequences representing specific pattern of the same highlight that are regarded as specialization of the highlight. Visual concepts for each highlight specialization are automatically obtained as the centers of these clusters.

Extraction of visual features is performed on MPEG videos, using both the compressed and uncompressed domain data. The MPEG motion vectors, that are used to

[1] http://dmag.upf.edu/ontologies/mpeg7ontos

calculate indexes of camera motion direction and intensity are extracted from the P and B frames. All the other visual features are extracted from the decompressed MPEG frames. In particular these features are the playfield shape, playfield lines and players blobs. From all these low-level features some higher level features are derived. In particular the playfield zone framed is recognized using naive Bayes classifiers that use particular shapes of the playfield region, the position of the playfield corner, the midfield line position and the orientation of the playfield lines; twelve different playfield zones that cover all the playfield are recognized. A thorough description of this process can be found in our previous work [2]. Combining the recognized playfield zone with the estimation of the number of players of each blob, and the blob position, the number of players in the upper and lower part of the playfield are obtained.

The visual features used to describe visual concepts within the pictorially enriched ontology and to perform the annotation of unknown sequences are: *i)* the playfield area, *ii)* the number of players in the upper part of the playfield, *iii)* the number of players in the lower part of the playfield, *iv)* the motion intensity, *v)* the motion direction, *vi)* motion acceleration.

The first step of the pictorially enriched ontology creation is to define for each clip a feature vector $V$ containing 6 distinct components. Each component is a vector $U$ that contains the sequence of values of each visual feature. The length of feature vectors $U$ may be different in different clips, depending on the duration and content of the clips. Vectors $U$ are quantized, and smoothed to eliminate possible outliers. Then the clustering process groups the clips of the representative set according to their visual features. We have employed the fuzzy c-means (FCM) clustering algorithm to take into account the fact that a clip could belong to a cluster, still being similar to clips of different clusters. The maximum number of clusters for each highlight has been heuristically set to 10. The distance between two different clips has been computed as the sum of all the normalized Needleman-Wunch edit distances between the $U$ components of the

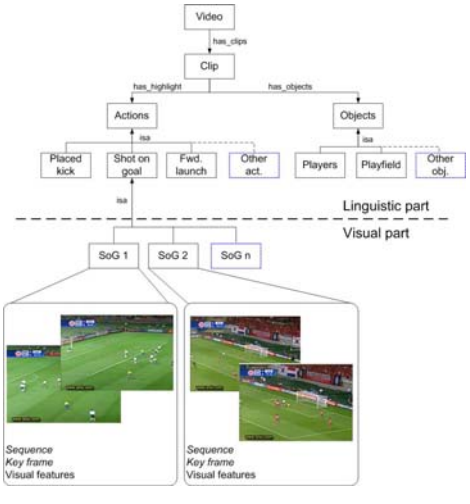


**Fig. 1.** Pictorially enriched ontology (partial view)

feature vector $V$ of the clips, to take into account the differences in the duration and the temporal changes of the features values. Performance evaluation of the generation of pictorially enriched ontology has been analyzed in our previous work [15]. The PE Ontology contains both soccer domain description and video domain description. They are related by the "visual" extension of the ontology through the exploitation of mid-level features assigned to each clip. In Fig. 2 the browsing interface of a PE Ontology is shown. It has to be noted that the user is able not only to browse with a single interface concepts related both to soccer and video domain but he can easily have at a glance the visual specifications of the linguistic concepts. When the user wants to see the different visual specifications of the linguistic concept "Shot on Goal", he can simply select the concept and the interface provides the clips that represent that concept. Moreover a cluster view of similar visual concepts related to a linguistic concept is provided.

## 4 Automatic Video Annotation Using Enriched Ontologies

To annotate the content of a video, in terms of highlights, two problems have to be solved: the detection of the part of the video where the highlight is, and the recognition of the highlight. The pictorially enriched ontology created with the process described in Sect. 3 can be used effectively to perform automatic video annotation with higher level concepts that describe what is occurring in the video clips. This is done by selecting clips that are to be annotated in the video, and checking the similarity of the clip content with the visual prototypes included in the ontology. If similarity is assessed with a

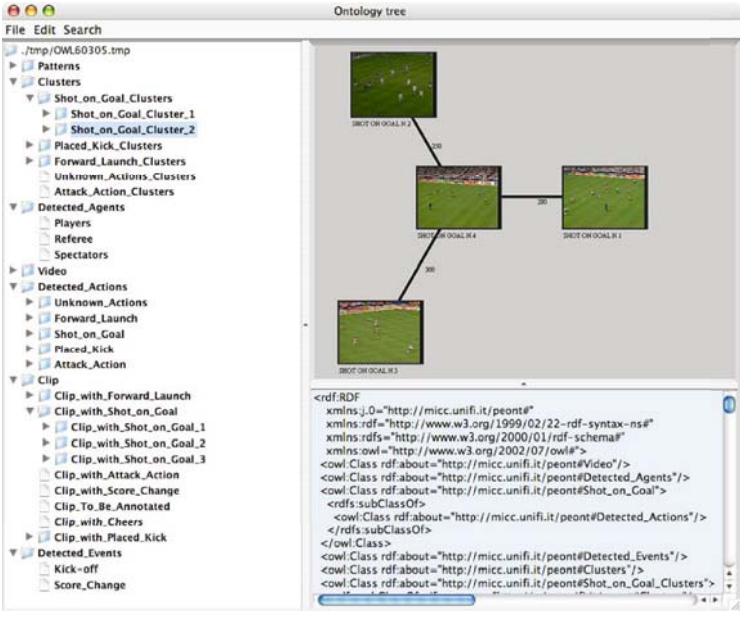


**Fig. 2.** A visualization of clips cluster for visual concept "Shot on Goal" within the PE Ontology Browser

particular visual concept then also higher level concepts in the ontology hierarchy, that are linked to the visual concept, are associated with the clip, resulting in a more complete annotation of the video content. The proposed annotation process is composed of two algorithms; a detailed description is provided in [16]. The first one selects the clips that are to be annotated from video sequences, such as shots or scenes automatically recognized or such as manual selections of clips, checking if they could contain some highlights; it is designed to be faster than performing an exhaustive analysis of all the clips that may be obtained within a sequence, partitioning it in sub-sequences. The second algorithm performs the annotation of the clips selected by the first algorithm.

The clip annotation algorithm is composed of two steps. In the first one an initial classification is performed evaluating the distance between visual prototypes and each clip. A clip is classified as an highlight type if its distance from a visual prototype is lesser than a computed threshold. In this step a special class (*Unknown action*) is created within the ontology, to hold all the clips that could not be classified by the algorithm. After each clip processing a FCM clustering is performed to re-evaluate the visual prototypes of the highlight. The second step analyzes each clip classified as *Unknown action*. A clip is classified as an highlight type if enough clips of that highlight type have a distance from the clip that is lesser than a computed threshold. If a clip is classified as an highlight type then FCM clustering is performed to re-evaluate the visual prototypes of this highlight.

At the end of the second step of the algorithm it is possible that some clips are still classified as types of *Unknown action*. These clips can be classified at later stage when other clips add more knowledge to the ontology, defining more visual prototypes or refining the clip classification according to the existing prototypes. The FCM clustering of clips annotated as *Unknown action* is performed to ease the manual annotation, allowing a user to annotate a whole cluster of clips. Among the clips that are classified as *Unknown action* there may be clips that do not contain an highlight, but that were selected by Alg. 1 as candidates to contain an highlight.

## 5 Querying the Ontology

Once videos are classified and annotated using the PE ontology it is possible to refine annotation by mean of reasoning on the ontology. In order to do this we have identified some “patterns” in the soccer video sequences in terms of series of detected actions and events. Analyzing the broadcasted video sequences we can notice, for instance, that if an attack action leads to a scored goal, cheers from spectators and superimposed text with score change are shown after the goal. We can identify a “pattern” for scored goal that contains possible combinations of detected actions and events and define a formal description of this pattern within the ontology by mean of conditions on class properties.

For instance the subclass *Video with scored goal*, which identify a sequence containing a scored goal, can be defined as *Video* that contains:

- Forward Launch Action followed by Shot on Goal Action followed by Cheers Event followed by Score Change Event or

- Placed Kick Action followed by Cheers Event followed by Score Change Event or
- Shot on Goal Action followed by Cheers Event followed by Score Change Event.

The reasoner evaluates the inferred types of each instance of the *Clip* and *Video* classes classifying them as a type of the proper subclass according to the detected actions or pattern they contain. For example a clip is classified as *Clip with Attack Action* by the reasoner if the *has highlight* property is related to subclass *Attack Action* of the *Detected Action* class. A video sequence is classified as *Video with scored goal* by the reasoner if the ordered sequence of clips contained corresponds to the pattern expressed by the *Patterns* class conditions.

The inferred type computation performed by the reasoner results in an enhanced annotation of video sequences and clips and allow to retrieve content performing complex queries on the ontology.

The PE Ontology structure and the visual prototype definition together with the enhanced annotation generated by the inference engine allow the user to easily perform semantic queries in order to retrieve video contents. For instance is possible to retrieve a video sequence asking for all the sequences that contain at least one placed kick, or for all the sequences that end with a forward launch or all the sequences with a scored goal that start with an attack action. Moreover, visual prototypes defined in the ontology can be used to refine retrieval using visual concepts. The system can retrieve all the video containing user-defined sequences of highlights and on each of them conditions of similarity to specific visual concept can be imposed. For the query user interface implementation we have used RACER [17] as description logic (DL) reasoner and nRQL [18] as query language.

A simple query interface that allows retrieval of video sequences composed by different video highlights, i.e. allowing retrieval of user defined patterns of actions has been integrated with the ontology browsing interface. Complexity of native nRQL query expression is transparent since the query interface is expressed in natural language (Fig. 3) and dynamically translates the user requests to formal nRQL queries.

## 6 Experimental Results

The proposed algorithms that perform automatic annotation, shown in Sect. 4, have been tested on MPEG-2 soccer videos from World Championship 2002, European Championship 2000 and 2004, recorded at 25 frame per second and with the resolution of $720 \times 576$ (PAL standard).

A set of representative sequences for three of the most important soccer highlights, namely shot on goal, placed kicks and forward launch, have been selected in order to create the pictorially enriched ontology. In particular 68 clips were manually annotated and selected (35 shots on goal, 16 forward launches and 17 placed kicks). The ontology creation process has been performed using this training set, obtaining 5 visual prototypes for shot on goal, 4 for forward launch and 3 for placed kick. Using this pictorially enriched ontology we have performed automatic video annotation on a different set of 242 clips that were automatically selected (85 shots on goal, 42 forward launches, 43 placed kicks and 72 that did not contain any highlight), using the process described

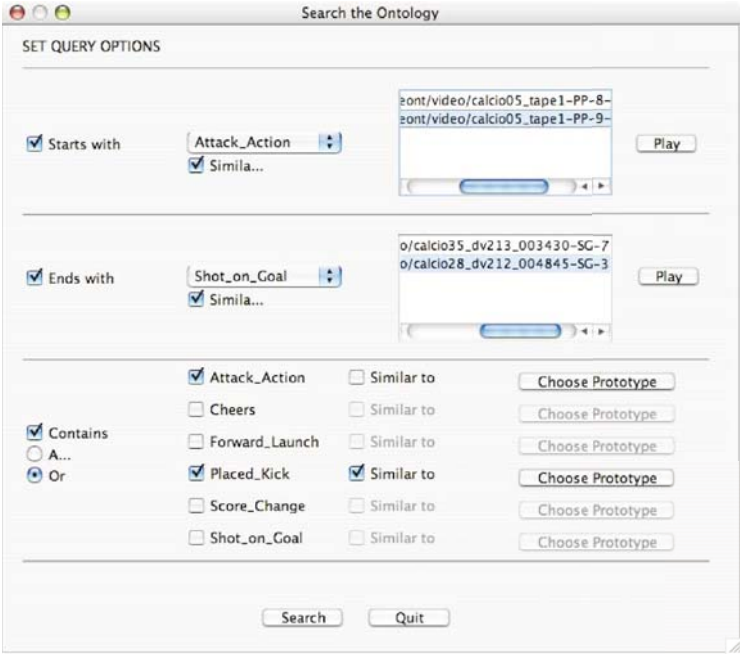


**Fig. 3.** PE Ontology query interface: the user is querying a sequence that starts with a attack action an finishes with a shot on goal. It is required that both actions are visually similar to a certain video clip. Moreover the video should contain any type of attack action or a placed kick visually similar to a model required by the user.

**Table 1.** Precision and recall of highlights/no highlights detection in video clips

| | Precision | Recall |
|---|---|---|
| Action | 100% | 59% |
| No action | 85% | 100% |

in Sect. 4. Table 1 reports precision and recall figures for the clip selection algorithm (Alg. 1), using the second test set. Table 2 reports precision and recall figures of Alg. 2 for the clips that were selected by Alg. 1 as candidates to contain highlights.

The goal of the clip selection algorithm is to detect all the clips that could contain possible highlights. To this end the conditions used to select a clip are loose enough to avoid misses of "Action" clips, while maintaining a relatively high precision figure of "No action" clips, as shown in Table 1. It has to be noted that at the end of this algorithm no clip has been annotated and inserted in the ontology, yet.

In the second table we have reported the percentage of clips that remained classified as *Unknown action* instead of reporting it in the Miss column because this kind of error may be corrected at a later stage, when more clips are fed to the system as described in Sect. 4. Anyway the figure of the clips classified as *Unknown action* has been taken into account to evaluate the recall performance. The algorithm aims to obtain the highest

**Table 2.** Precision and recall of highlights classification

| Highlight | Miss | False | Unknown | Precision | Recall |
|---|---|---|---|---|---|
| Shot on goal | 5% | 16% | 21% | 82% | 74% |
| Placed kick | 9% | 9% | 18% | 89% | 73% |
| Fwd. launch | 10% | 10% | 30% | 86% | 60% |

values of precisions at the expense of recall since it is more convenient to classify a clip as *Unknown action* if there is some uncertainty rather then to risk that it becomes a prototype for a wrong visual concept. In fact the FCM clustering performed at the end of each classification step, in some cases, may select the wrong clip as cluster center and then as visual prototype of the ontology, even if this did not happen in our experiments.

The results reported in Table 2 are obtained from the annotation process of the clips selected by the clip selection algorithm; some of the false detections are then due to clips that were selected as possible "Action" clips, but that actually did not contain any highlight. In fact some slow play close to the goal box area were wrongly classified as placed kick, due to the similarity with the initial part of the placed kick, in terms of motion and playfield area framed. Other false detections may be due to wrong highlight classification: forward launches and shot on goals may be confused since both actions have similar behaviour in motion intensity and direction. Placed kicks have usually higher length then shot on goals, due to an initial part containing almost no motion where the players get prepared for the kick. In some cases the broadcasted video that we used in the experiments does not include this part of the placed kick, and thus they have a behaviour in terms of playfield area, motion and length that is very similar to that of shots on goal. Inspection of the clusters composed by clips annotated as *Unknown action* reported similar precision values of the annotated clips, thus a user may confidently annotate an entire cluster manually, simply inspecting the visual prototype.

## 7 Conclusions

The novelty of this paper is the presentation of pictorially enriched ontologies based both on linguistic and visual concepts and the implementation of two algorithms that perform automatic annotation of soccer video based on these extended ontologies.

With the proposed method annotation is performed automatically associating occurrences of events or entities to higher level concepts by checking their proximity to visual concepts that are hierarchically linked to higher level semantics, and applying reasoning to the ontology it is possible to exploit the domain knowledge and perform higher-level semantic annotation.

Differences between MPEG-7 and RDF/OWL have been presented and advantages of usage of knowledge description languages in automatic video annotation have been shown. Proper definition of specific domain ontologies, together with video and clip ontology describing low and mid level features, can improve the annotation process and provide user friendly visualization and query interfaces for video content browsing and retrieval.

Experimental results have been presented for typical soccer highlights in terms of precision and recall, showing that with pictorially enriched ontologies it is possible to perform automatic clips annotation up to the level of detail of pattern specification.

Our future work will deal with the improvement of the visual features set, the metrics and distances used in the ontology creation and in the annotation process, the improvement of the proposed interfaces for browsing and querying the pictorially enriched ontologies.

*Acknowledgment.* 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. Leonardi, R., Migliorati, P.: Semantic indexing of multimedia documents. IEEE Multimedia **9**(2) (2002) 44–51
2. Assfalg, J., Bertini, M., Colombo, C., Bimbo, A.D., Nunziati, W.: Semantic annotation of soccer videos: automatic highlights identification. Computer Vision and Image Understanding **92**(2-3) (2003) 285–305
3. X.Yu, Xu, C., Leung, H., Tian, Q., Tang, Q., Wan, K.W.: Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In: ACM Multimedia 2003. Volume 3., Berkeley, CA (USA) (2003) 11–20
4. Reidsma, D., Kuper, J., Declerck, T., Saggion, H., Cunningham, H.: Cross document ontology based information extraction for multimedia retrieval. In: Supplementary proc. of the ICCS03, Dresden (2003)
5. Mezaris, V., Kompatsiaris, I., Boulgouris, N., Strintzis, M.: Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. IEEE Transactions on Circuits and Systems for Video Technology **14**(5) (2004) 606–621
6. Jaimes, A., Tseng, B., Smith, J.: Modal keywords, ontologies, and reasoning for video understanding. In: Int'l Conference on Image and Video Retrieval (CIVR 2003). (2003)
7. Jaimes, A., Smith, J.: Semi-automatic, data-driven construction of multimedia ontologies. In: Proc. of IEEE Int'l Conference on Multimedia & Expo. (2003)
8. Benitez, A., Chang, S.F.: Automatic multimedia knowledge discovery, summarization and evaluation. IEEE Transactions on Multimedia, Submitted (2003)
9. Strintzis, J., Bloehdorn, S., Handschuh, S., Staab, S., Simou, N., Tzouvaras, V., Petridis, K., Kompatsiaris, I., Avrithis, Y.: Knowledge representation for semantic multimedia content analysis and reasoning. In: European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology. (2004)
10. Dasiopoulou, S., Mezaris, V., Kompatsiaris, I., Papastathis, V.K., Strintzis, M.G.: Knowledge-assisted semantic video object detection. IEEE Transactions on Circuits and Systems for Video Technology **15**(10) (2005) 1210–1224
11. Hunter, J.: Adding multimedia to the semantic web: Building an MPEG-7 ontology. In: The First Semantic Web Working Symposium, SWWS01, Stanford University, California, USA (2001)
12. Hunter, J.: An RDF schema/DAML+OIL representation of MPEG-7 semantics. Technical Report MPEG Document: ISO/IEC JTC1/SC29/WG11 W7807, ISO/IEC (2001)
13. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support for ontology-based video retrieval applications. In: Proc. Image and Video Retrieval: Third International Conference, CIVR 2004, Image and Video Retrieval: Third International Conference, CIVR 2004, Lecture Notes in Computer Science 3115 Springer (2004)

14. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability of OWL with the MPEG-7 MDS. Technical report, Technical University of Crete / Laboratory of Distributed Multimedia Information Systems and Applications (TUC/MUSIC) (2004)
15. Bertini, M., Cucchiara, R., Del Bimbo, A., Torniai, C.: Video annotation with pictorially enriched ontologies. In: Proc. of IEEE Int'l Conference on Multimedia & Expo. (2005)
16. Bertini, M., Del Bimbo, A., Torniai, C.: Enhanced ontologies for video annotation and retrieval. In: Proceedings of ACM MIR. (2005)
17. Haarslev, V., Möller, R.: Description of the racer system and its applications. In: Proceedings International Workshop on Description Logics (DL-2001), Stanford, USA, 1.-3. August. (2001) 131–141
18. Haarslev, V., Möller, R., Wessel, M.: Querying the semantic web with racer + nrql. In: Proceedings of the KI-2004 International Workshop on Applications of Description Logics (ADL'04), Ulm, Germany, September 24. (2004)