Enhanced Ontologies for Video Annotation and Retrieval

Marco Bertini, Alberto Del Bimbo, Carlo Torniai
Università di Firenze - Italy
Via S. Marta, 3 - 50139 Firenze
bertini@dsi.unifi.it, delbimbo@dsi.unifi.it, torniai@dsi.unifi.it

ABSTRACT
A typical way to perform video annotation requires to classify video elements (e.g., events and objects) according to some pre-defined ontology of the video content domain. Ontologies are defined by establishing relationships between linguistic terms that specify domain concepts at different abstraction levels. However, although linguistic terms are appropriate to distinguish event and object categories, they are inadequate when they must describe specific or complex patterns of events or video entities. Instead, in these cases, pattern specifications can be better expressed using visual prototypes, either images or video clips, that capture the essence of the event or entity. Therefore enhanced ontologies, that include both visual and linguistic concepts, can be useful to support video annotation up to the level of detail of pattern specification.

This paper presents algorithms and techniques that employ enriched ontologies for video annotation and retrieval, and discusses a solution for their implementation for the soccer video domain. An unsupervised clustering method is proposed in order to create pictorially enriched ontologies by defining visual prototypes that represent specific patterns of highlights and adding them as visual concepts to the ontology.

Two algorithms that use pictorially enriched ontologies to perform automatic soccer video annotation are proposed and results for typical highlights are presented. Annotation is performed associating occurrences of events, or entities, to higher level concepts by checking their similarity to visual concepts that are hierarchically linked to higher level semantics, using a dynamic programming approach.

Usage of reasoning on the ontology is shown, to perform higher-level annotation of the clips using the domain knowledge and to create complex queries that comprise visual prototypes of actions, their temporal evolution and relations.

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MIR'05, November 10–11, 2005, Singapore.
Copyright 2005 ACM 1-59593-244-5/05/0011 ...$5.00.

General Terms
Algorithms

Keywords
Automatic video annotation, clustering, ontologies, RDF, OWL, Dynamic programming

1. INTRODUCTION AND PREVIOUS WORK
Ontologies are formal, explicit specifications of a domain knowledge: they consist of concepts, concept properties, and relationships between concepts and are typically represented using linguistic terms.

In the last years several standard description languages for the expression of concepts and relationships in domain ontologies have been defined: Resource Description Framework (RDF) [7], Resource Description Framework Schema (RDFS), Web Ontology Language (OWL) [8] and the XML Schema in MPEG-7. Using these languages metadata can be tailored to specific domains and purposes, yet still remaining interoperable and capable of being accessed 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, more recently and effectively, 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-level features. 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 [15] where MPEG motion vectors, playfield shape and players position have been used with Hidden Markov Models to detect soccer highlights. Ekin et al., in [10], have assumed that the presence of soccer highlights can be inferred from the occurrence of one or several slow motion shots and from the presence of shots where the referee and/or the goal post is framed. In [1] 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. [19] 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 [17]. In this paper, the annotation engine makes use of reasoning algorithms to automatically create a semantic annotation of soccer video sources. In [16], 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. However, 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. Consider for example the many different ways in which an attack action can occur in soccer. We can easily distinguish several different patterns that differ each other by the playfield zone, the number of players involved, the player’s motion direction, the speed, etc. Each of these patterns represents a specific type of attack action that could be expressed in linguistic terms only with a complex sentence, explaining the way in which the event has developed.

The possibility of extending linguistic ontologies with multimedia ontologies, has been suggested in [14] 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 [13]: text information available in videos and visual features are extracted and manually assigned to concepts, properties, or relationships in the ontology. In [11] 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 [18] 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 [9]. 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. Think for example, in the soccer domain, of a TV sport program editor that is interested in selecting similar actions, e.g. Beckham’s bent free kicks that differ from other players’ free kicks. In this case, it is important that the highlight patterns that share similar spatio-temporal behaviours are clustered and described with one single concept that is a specialization of the free kick term in the linguistic ontology. These requirements motivate the possibility that events that share the same patterns are 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 the plethora of visual events 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. In order to extend a linguistic ontology with visual information a set of representative sequences containing highlights described in the linguistic ontology is selected, visual features are extracted from the sequences and an supervised clustering process is performed. Clusters of sequences representing specific patterns of the same highlight are generated. Centers of these clusters are regarded as visual concepts of each highlight pattern and are added to the linguistic ontology as specialization of the linguistic concept describing the highlight. Clusters centers and members are used also to perform automatic annotation of video clips. This process creates the pictorially enriched ontologies assigning multimedia objects to concepts and integrating the semantics described by the linguistic terms. Reasoning on the ontology, using description logic, allows to add 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:

- the unification in the same ontology both of high level linguistic concepts and lower-mid level concepts (typically representing patterns of actions, or special occurrences of entities, that are difficult to be represented in linguistic terms but are better expressed by visual data); while higher concepts are related each other manually, lower level concepts can be created automatically by appropriate clustering of low-mid level features;
- the capability to associate automatically occurrences of events or entities to higher level concepts checking their proximity to visual concepts that are hierarchically linked to higher level semantics.

The paper is organized as follows: 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. 2. Two algorithms that use the enriched ontology to perform automatic annotation are presented in Sect. 3. In Sect. 4 is shown how ontology-based reasoning adds a more refined annotation to the videos, allowing the retrieval of video content by mean of complex queries on the ontology. In Sect. 5 we discuss the preliminary results of the proposed system applied to soccer videos annotation. Finally, in Sect. 6 we provide conclusions and some future works.

2. 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 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. Playfield shape is segmented using color histograms and grass color information. This shape is refined applying a processing chain of K-fill, flood fill and the erosion and dilatation morphological operations, and is represented as a polygon. The playfield lines are extracted from the edge image of the playfield region using a stick growing algorithm; these lines are joined together when they are close and collinear. Lines length and color information of the area around the lines are used to further refine the playfield lines recognition. Players blobs are segmented from the playfield using color differencing and morphological operators. Width/height ratio of the blobs bounding boxes and blob/box area ratio are used to refine their detection and perform an estimation of the number of players.

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 [1]. 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:

- the playfield area;
- the number of players in the upper part of the playfield;
- the number of players in the lower part of the playfield;
- the motion intensity;
- the motion direction;
- the 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 ([6]) 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 distances between the $U$ components of the feature vector $V$ of the clips, to take into account the differences in the duration and the temporal changes of the features values. This distance is a generalization of the Levenshtein edit distance and has been used since the cost of character substitutions is an arbitrary distance function. In our case the cost is used to weight differently the differences in the motion intensity. The normalization is used in order to better discriminate differences between short and long sequences and is performed dividing the Needleman-Wunch distance by the length of the shorter sequence. Performance evaluation of the generation of pictorially enriched ontology has been analyzed in our previous work [5].

![Figure 1: Pictorially enriched ontology (partial view)](image-url)
3. 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. 2 can be used effectively to perform automatic video annotation with higher level concepts that describe what is occurring in the video clips. This is made 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 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. The first one, shown in Alg. 1, 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.

In the clip selection algorithm (Alg. 1) the distance $\text{lcs}(a, b)$ is the average length of the longest common subsequence between all the $U$ components of the feature vector $V$ of clips $b$ and $c$; this distance has been used since it easily finds the start and end positions of the common parts of the $U$ components. A dynamic programming approach for the computation of this distance has been followed, and the complexity of the solution is $O(mn)$, where $m$ and $n$ are the length of the $U$ component of the prototype and of the $U$ component of the video sequence being analyzed. $\tau_{\text{es}}$ is computed as half the length of the prototype. $\tau_{\text{ran}}$ is computed as the length of $\text{lcs}$ plus 10%. The combination of the two conditions means that we will select for further analysis those clips that have a minimum similarity w.r.t. a prototype, and that this similar portion has to be compact enough to resemble to the temporal evolution of the prototype. The range in which the clips are selected is calculated from the average positions of the $\text{lcs}$ of the $U$ vectors; this range is enlarged by 25% of the length of the prototype, to take into account the fact that the evaluation of the $\text{lcs}$ allows only exact match of the strings; moreover using this range value it is possible to handle the case of multiple matching prototypes. The clips selected by this algorithm are used in the annotation algorithm.

$\tau_{\text{es}}$ is calculated in order not to be conservative, since the $\text{lcs}$ does not take into account substitutions, opposite to the Needleman-Wunch distance used in the annotation algorithm.

The clip annotation algorithm is composed of two steps, and is shown in Alg. 2. 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.

$\text{d}(b, c)$ is the sum of all the normalized Needleman-Wunch distances between the $U$ components of the feature vector $V$ of the clips $b$ and $c$; the Needleman-Wunch distance is evaluated using a dynamic programming algorithm, and particularly suited for this task since it is the most used methods for global alignments of strings. $\tau_{1}$ is computed as half of the minimum of the distances between all the visual prototypes in the ontology; $\tau_{2}$ is computed as the average of the radius of all the clusters of one specific highlight; $X$ is an highlight within the ontology, with $X = 1...\text{number of highlights classes}$; $k$ is computed as the average of the number of clips of each cluster.

$\tau_{1}$ is calculated in order to avoid misclassification due to the possible lack of knowledge contained in the ontology at this initial stage. In fact we have to take in to account that the ontology creation process could have been performed using a training set that does not include representative visual concepts of certain types of highlights. $\tau_{2}$ is calculated to be less conservative because it is evaluated at each step and uses all the knowledge that has been added to the ontology by the annotation process. At the end of the

Algorithm 1 Video clip selection using the pictorially enriched ontology

for each video sequence
    perform visual feature extraction
    for each visual prototype
        evaluate $\tau_{\text{es}}$
        calculate distance and range $\text{lcs}(\text{video sequence, visual prototype})$
        if $\text{lcs} > \tau_{\text{es}} \land \text{range} < \tau_{\text{ran}}$ then
            select sub-sequence in range as clip for annotation

Algorithm 2 Video clip annotation using the pictorially enriched ontology

First step
for each clip
    evaluate $\tau_{1}$
    perform visual feature extraction
    for each visual prototype
        calculate distance $\text{d}(\text{clip, visual prototype})$
        if $\text{d} < \tau_{1}$ then
            classify clip according to the highlight of visual prototype
        else
            classify clip as Unknown action
            perform FCM clustering on all the highlight clusters

Second step
repeat
    evaluate $\tau_{2}$
    for each clip classified as Unknown action
        calculate distance $\text{d}(\text{clip, classified clips})$
        let $n_{X} = \text{number of clips with } d < \tau_{2}$, classified as $X$
        if $M = \max(n_{X}) > k$ then
            classify clip as $X$
            perform FCM clustering on all the highlight clusters
            perform FCM clustering on the Unknown action’s clusters
        until number of unclassified clips changes

92
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.

4. REASONING ON 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.

In Fig. 2 a simplified class hierarchy of the PE Ontology is shown. The *Clusters* class contains the clusters of annotated clips. Subclasses of Video and Clip have been defined according to the their highlights. For the Video Class additional subclasses that contain a specific pattern have been defined, i.e Video with scored goal and Video with missed goal.

For instance the subclass Video with scored goal, which identifies 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 score change event is detected according to the OCR procedure described in [4], while the crowd detector used to classify Cheers Event is the one described in [2]. The following sample (Listing 1) shows the OWL code that express the above conditions on the subclass. In particular the OWL code between line 3 and 36 describes the first possible path that leads to the scored goal, lines between 37 and 60 describe the second path and lines between 61 and 84 the third one. In every path the *followed by* relation has been translated in the *has clip x* property with *x* representing the number of clip in the sequence ordered by time.

**Listing 1: Example of Video with scored goal class expressed in OWL**

```xml
<owl:Class rdf:ID="Video_with_Scored_Goal">
  <owl:equivalentClass>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class/>
        <owl:intersectionOf rdf:parseType="Collection">
          <owl:Restriction>
            <owl:onProperty>
```
corded at 25 frame per second and with the resolution of 720×576 (PAL standard).

In Sect. 3, have been tested on MPEG-2 soccer videos from World Championship 2002, European Championship 2000 and 2004, re-

taining visual concepts similar to the visual prototype Clip 1. 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 class conditions.

The inferred type computation performed by the reasoner results in an enhanced annotation of video sequences and clips and allows to retrieve content performing complex queries on the ontology. 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 retrieve sequences by visual concepts. Given one or more prototypes of an highlight the system can retrieve all the video sequences containing similar visual concepts. We have used RACER [11] as description logic (DL) reasoner and nRQL [12] as query language.

The following query example, expressed in nRQL, shows how to retrieve all the sequences containing visual concepts similar to the visual prototype Clip 1.

Listing 2: Example of nRQL query: retrieve all sequences containing visual concepts similar to the visual prototype Clip 1

```
RETRIEVE

(∀Z)

(AND (∀X | Clip_1 | | has_prototype |)

(∀Y | Video | | has_members |)

(∀Z | Video | | has_clips |)

```

In Listing 2, initially the cluster X containing the visual concept (line 3) and all the cluster members Y are identified (line 4), then all the sequences Z containing similar clips (i.e. the Y members) are retrieved (line 5 and 6).

5. EXPERIMENTAL RESULTS

The proposed algorithms that perform automatic annotation, shown in Sect. 3, 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 720×576 (PAL standard).

A set of representative sequences for three of the most important soccer highlights, namely shot on goal, placed kicks and for-
ward 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 in Sect. 3. 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. 3. 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 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 than 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 may be 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. In a few cases attack actions that finished with a ball kicked between the corner and the goal post were wrongly classified as shot on goal because the motion features are not precise enough to discriminate these two cases. 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. This happens the most when a forward launch action is longer than usual and thus the normalization of the Needleman-Wunch distance becomes less discriminating. 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.

### 6. CONCLUSIONS

This paper presents pictorially enriched ontologies based both on linguistic and visual concepts and the implementation of solutions for video annotation and retrieval based on these extended ontologies.

In order to create pictorially enriched ontologies an unsupervised clustering method has been proposed. The clustering process defines visual prototypes representing specific patterns of highlights and adds them as visual concepts to the linguistic ontology. Results for the two proposed algorithms for automatic clip selection and annotation of soccer video using pictorially enriched ontologies have been presented in terms of precision and recall. Experiments have shown that with pictorially enriched ontologies it is possible to perform automatic clips annotation and retrieval up to the level of detail of pattern specification.

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. Furthermore we have defined patterns in terms of series of detected actions and events that allow to exploit the domain knowledge and perform higher-level semantic annotation applying reasoning to the ontology.

Our future work will deal with the improvement of the visual features set, the optimization of metrics and distances used in the ontology creation and in the annotation process, the definition of more actions patterns for the annotation process refinement, the investigation of the usefulness of synthetic visual prototypes and the generalization and extension of pictorially enriched ontology to other domains.

### 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).

### 7. REFERENCES


### Table 1: Precision and recall of highlights/no highlights detection in video clips

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>100%</td>
<td>59%</td>
</tr>
<tr>
<td>No action</td>
<td>83%</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Table 2: Precision and recall of highlights classification

<table>
<thead>
<tr>
<th>Highlight</th>
<th>Miss</th>
<th>False</th>
<th>Unknown</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shot on goal</td>
<td>5%</td>
<td>16%</td>
<td>21%</td>
<td>82%</td>
<td>74%</td>
</tr>
<tr>
<td>Placed kick</td>
<td>9%</td>
<td>9%</td>
<td>18%</td>
<td>89%</td>
<td>73%</td>
</tr>
<tr>
<td>Fwd. launch</td>
<td>10%</td>
<td>10%</td>
<td>30%</td>
<td>86%</td>
<td>60%</td>
</tr>
</tbody>
</table>


