# Automatic Video Annotation using Ontologies Extended with Visual Information

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

Classifying video elements according to some pre-defined ontology of the video content domain is a typical way to perform video annotation. 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 patterns of events or video entities. Instead, in these cases, pattern specifications can be better expressed through visual prototypes that capture the essence of the event or entity. Therefore *pictorially enriched 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 pictorially enriched ontologies and discusses a solution for their implementation for the soccer video domain. An unsupervised clustering method is proposed in order to create the enriched ontologies by defining visual prototypes representing specific patterns of highlights and adding them as visual concepts to the ontology.

An algorithm that uses pictorially enriched ontologies to perform automatic soccer video annotation is proposed and results for typical highlights are presented. Annotation is performed 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.

**Categories and Subject Descriptors:** H.3.7 [Information Storage and Retrieval]: Digital Libraries; H.2.4 [Systems]: Multimedia databases
**General Terms:** Algorithms
**Keywords:** Automatic video annotation, clustering, ontologies, RDF

## 1. INTRODUCTION

Ontologies are formal, explicit specifications of a domain knowledge: they consist of concepts, concept properties, and relationships between concepts that 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) , Resource Description Framework Schema (RDFS), Web Ontology Language (OWL) 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-mid 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 [8] where MPEG motion vectors, playfield shape and players position have been used with Hidden Markov Models to detect soccer highlights. Ekin et al., in [5], 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. [12] 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 recognised 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 [10]. In this paper, the annotation engine makes use of reasoning algorithms to automatically create a semantic annotation of soccer video sources. In [9], 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 [7] 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


*MM'05*, November 6–11, 2005, Singapore.

presented to automatically classify keywords from speech recognition, queries or related text into these categories. Multimedia ontologies are constructed manually in [6]: text information available in videos and visual features are extracted and manually assigned to concepts, properties, or relationships in the ontology. In [2] 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 [11] 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.

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 organised 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 unsupervised 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. This process creates the pictorially enriched ontologies assigning multimedia objects to concepts and integrating the semantics described by the linguistic terms. 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. An algorithm that uses the enriched ontology to perform automatic annotation is presented in Sect. 3. In Sect. 4 we discuss the preliminary results of the proposed system applied to soccer videos annotation. Finally, in Sect. 5 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 the 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 RDF 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 dilation 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 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 ([4]) 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 [3].

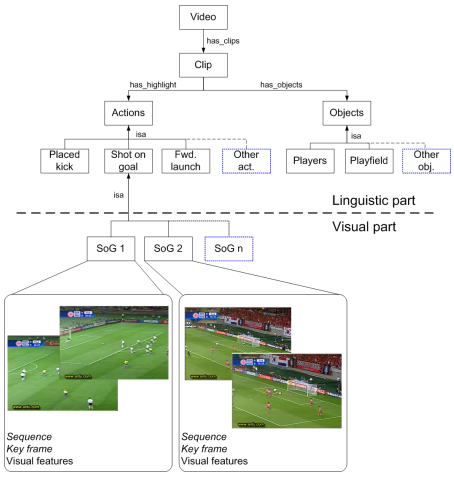


**Figure 1: Pictorially enriched ontology (partial view)**

# 3. AUTOMATIC VIDEO ANNOTATION USING PICTORIALLY ENRICHED ONTOLOGY

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 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 algorithm is shown in Alg. 1. The 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 (*Not recognized*) 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 *Not recognized*. 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.

In Algorithm 1 the distance $d(b, c)$ is the sum of all the normalized Needleman-Wunch distances between the $U$ components of the feature vector $V$ of the clips $b$ and $c$; $\tau_1$ 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

**Algorithm 1** Video clip annotation using the pictorially enriched ontology

**First step**
- evaluate $\tau_1$
- for each clip
  - perform visual feature extraction
  - for each visual prototype
    - calculate distance $d(clip, visual\ prototype)$
    - if $d < \tau_1$ then
      - classify clip according to the highlight of visual prototype
    - else
      - classify clip as *Not recognized*
  - perform FCM clustering on all the highlight clusters

**Second step**
- repeat
  - evaluate $\tau_2$
  - for each clip classified as *Not recognized*
    - calculate distance $d(clip, classified\ clips)$
    - let $n_X$ = 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 *Not recognized's* clusters
- until number of unclassified clips changes

ontology, with $X = 1..$ 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 second step of the algorithm it is possible that some clips are still classified as types of *Not recognized*. 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 *Not recognized* is performed to ease the manual annotation, allowing a user to annotate a whole cluster of clips.

# 4. EXPERIMENTAL RESULTS

The proposed algorithm has 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 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 170 clips that were manually selected (85 shots on goal, 42 forward launches and 43 placed kicks), using the process described in Sect. 3. Table 1 reports precision and recall figures for the three highlights.

*Not recognized* clips cannot be associated with any visual prototypes using the current knowledge of the ontology. These clips may be annotated at a later stage, when more clips are fed to the system as described in 3. Anyway the figure of the clips classified as *Not recognized* has been taken into account to evaluate the recall performance, considering the clips as misses. 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 *Not recognized* 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.

Results reported in the table show that 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 *Not recognized* reported similar precision values of the annotated clips, thus a user may confidently annotate an entire cluster manually.

| Highlight | Miss | False | Not Recog. | Precision | Recall |
|---|---|---|---|---|---|
| Shot on goal | 5% | 11% | 21% | 88% | 74% |
| Placed kick | 9% | 0% | 18% | 100% | 73% |
| Fwd. launch | 10% | 10% | 30% | 86% | 60% |

**Table 1: Precision and recall of clips classification**

# 5. CONCLUSIONS

The novelty of this this paper is the definition of pictorially enriched ontologies based both on linguistic and visual concepts and the implementation of an algorithm that performs automatic annotation of soccer video based on these extended ontologies.

An unsupervised clustering method has been proposed in order to create pictorially enriched ontologies by defining visual prototypes representing specific patterns of highlights and adding them as visual concepts to the linguistic ontology. Results for the proposed algorithm for automatic annotation of soccer video using pictorially enriched ontologies have been presented for typical highlights in terms of precision and recall. Experiments have shown that with pictorially enriched ontologies it is possible to perform automatic clips annotation 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.

Our future work will deal with the improvement of the visual features set, the optimization of metrics used in the ontology creation and in the annotation process, the creation 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).

# 6. REFERENCES

[1] J. Assfalg, M. Bertini, C. Colombo, A. D. Bimbo, and W. Nunziati. Semantic annotation of soccer videos: automatic highlights identification. *Computer Vision and Image Understanding*, 92(2-3):285–305, November-December 2003.

[2] A. Benitez and S.-F. Chang. Automatic multimedia knowledge discovery, summarization and evaluation. *IEEE Transactions on Multimedia, Submitted*, 2003.

[3] M. Bertini, R. Cucchiara, A. Del Bimbo, and C. Torniai. Video annotation with pictorially enriched ontologies. In *Proc. of IEEE Int'l Conference on Multimedia & Expo*, 2005.

[4] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[5] 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.

[6] A. Jaimes and J. Smith. Semi-automatic, data-driven construction of multimedia ontologies. In *Proc. of IEEE Int'l Conference on Multimedia & Expo*, 2003.

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

[8] R. Leonardi and P. Migliorati. Semantic indexing of multimedia documents. *IEEE Multimedia*, 9(2):44–51, April-June 2002.

[9] 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.

[10] D. Reidsma, J. Kuper, T. Declerck, H. Saggion, and H. Cunningham. Cross document ontology based information extraction for multimedia retrieval. In *Supplementary proceedings of the ICCS03*, Dresden, July 2003.

[11] 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.

[12] 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 2003.