# VIDEO ANNOTATION WITH PICTORIALLY ENRICHED ONTOLOGIES

*M. Bertini†, R. Cucchiara‡, A. Del Bimbo†, C. Torniai†*

†bertini,delbimbo,torniai@dsi.unifi.it - D.S.I. - Università di Firenze - Italy
‡rita.cucchiara@unimo.it -D.I.I. - Università di Modena e Reggio Emilia - Italy

**ABSTRACT**

Video annotation is typically performed by classifying video elements 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 patterns of events or video entities. Instead, in these cases, pattern specifications are better expressed through visual prototypes that capture the essence of the event or entity. *Pictorially enriched ontologies*, that include *visual concepts* together with linguistic keywords, are therefore needed to support video annotation up to the level of detail of pattern specification. This paper presents pictorially enriched ontologies and provide a solution for their implementation in the soccer video domain. The pictorially enriched ontology is used both to directly assign multimedia objects to concepts, providing a more meaningful definition than the linguistics terms, and to extend the initial knowledge of the domain, adding subclasses of highlights or new highlight classes that were not defined in the linguistic ontology. Automatic annotation of soccer clips up to the pattern specification level using a pictorially enriched ontology is discussed.

## 1. INTRODUCTION

Semantic annotation of video content is performed by using appropriate domain-specific ontologies that model the video content domain. Ontologies are formal, explicit specifications of the knowledge domain: they consist of concepts, concept properties, and relationships between concepts. Ontologies typically represent concepts by linguistic terms. However, also multimedia ontologies can be created, that assign multimedia objects to concepts.

Standard description languages for the expression of concepts and relationships in domain ontologies have been defined, like Resource Description Framework (RDF) [1], Resource Description Framework Schema (RDFS) and the XML Schema in MPEG-7. In this way metadata can be tailored to specific domains and purposes, yet still remaining interoperable and capable of being accessed by standard tools and search systems.

Semantic annotation is either performed manually, by 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-intermediate 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 sport video domain, it is possible to classify highlight events in different classes, like *shot on goal*, *counterattack*, *corner kick*, etc.

Examples of automatic semantic annotation systems have been presented recently, most of them in the application domain of sports video. Among these, in [2] MPEG motion vectors, playfield shape and players position have been used with Hidden Markov Models to detect soccer highlights. In [3], Ekin et al. 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 [4] 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. The ball trajectory has been used by Yu et al. [5] in order to detect the main actions like touching and passing and compute ball possession by 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 ontology-based representation of the domain. Domain specific linguistic ontology with multilingual lexicons, and possibility of cross document merging has instead been presented in [6]. In this paper, the annotation engine makes use of reasoning algorithms to automatically create a semantic annotation of soccer video sources. In [7], 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 patterns in which an attack action can occur in soccer. We can easily distinguish several different subclasses that differ each other by the playfield zone, the number of players involved, the player's motion direction, the speed. Each of these subclasses specifies a specific pattern of attack action that could be expressed in linguistic terms only with a complex sentence, explaining the way in which the event has developed. Despite of the difficulty of including pattern specifications into linguistic ontologies, classification at the pattern description level is mandatory, in many real operating contexts. Think for example, in the soccer domain, of a coach that is interested in the analysis of the ways in which the attack actions of his team have developed. 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 attack action term in the video 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 specifi-

*0-7803-9332-5/05/$20.00 ©2005 IEEE*

cations represented instead through visual concepts, can be both organized into new extended ontologies, that will be referred to 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. Visual concepts of pictorially enriched ontologies, like linguistic concepts, can be expressed in RDF, and therefore used in a search engine to perform content based retrieval from video databases or to provide video summaries. This paper discusses pictorially enriched ontologies and provide a solution for their implementation for soccer video automatic annotation of highlight patterns. The highlights detected by the annotation engine define the initial linguistic ontology. In order to distinguish specific patterns of the principal highlights additional visual features are added to the ontology. A clustering algorithm is used to create new subclasses of highlights representing specific patterns of the event and to group the clips within highlights subclasses according to their visual features. The visual concepts of the patterns of recognized highlights are automatically obtained as the centers of the clusters in which the video clip instances of the highlight are grouped. Once detected, visual concepts are added as prototypes in the ontology, to represent visually the appearance of the pattern category and integrate the semantics described by the linguistic terms. The ontology created is used both to directly assign multimedia objects to concepts and to extend the initial knowledge of the domain, adding subclasses of highlights or new highlight classes that were not defined in the linguistic ontology. Pictorially enriched ontologies are then used to support video annotation up to very specialized levels of pattern specification. The possibility of extending linguistic ontologies with multimedia ontologies, although with a different idea, has also been suggested in [8] to support video understanding. Differently from our contribution, the authors suggest to use *modal keywords*, i.e. keywords that represent perceptual concepts in the 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 ([9]): text information available in videos and visual features are extracted and manually assigned to concepts, properties, or relationships in the ontology. In [10] 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 linked to the domain ontology. In this paper an improvement to this approach is proposed, including visual features in the domain ontology and using a clustering algorithm that extends the domain ontology through visual features analysis. The paper is organized as follows: in Sect. 2 we present a prototype system for automatic semantic video annotation and discuss visual feature extraction. Creation of pictorially enriched ontologies for the representation of highlight patterns are discussed in Sect. 3. In Sect. 4 we discuss the preliminary results of the proposed system applied to soccer videos. Finally, in Sect. 5 we provide conclusions and some future works.

## 2. SOCCER HIGHLIGHT AUTOMATIC VIDEO ANNOTATION

The annotation system performs semantic annotation of MPEG videos, by detecting attack actions and placed kicks and whether or not they are terminated with a shot on goal. Highlights are detected by using a limited set of visual features that are extracted respectively: *i)* from the compressed domain: motion vectors (used to calculate indexes of camera motion direction and intensity); YUV color components (used to extract and evaluate the playfield shape that is framed); *ii)* from the uncompressed domain (uncompressed I and P frames): the ratio between the pixels of the players of the two teams (by exploiting the a-priori knowledge of team colors); the playfield lines filtered out on the basis of their length and orientation (used to recognize the playfield zone that is framed).

Frames are classified as close-, medium- and long-view, depending on the image-playfield ratio; long-view frames are further distinguished into left, central and right part of the playfield.

Evidences and inferences of highlights are computed for each MPEG GOP (typically 12 frames, about 1/2 second in PAL video standard). Four Bayes networks are used to predict highlights: two networks are used to predict (left, right) attack actions and two networks to predict (left, right) placed kicks. If the highlight is predicted, in the following 6 seconds (12 GOPs) the video is processed by two different Bayesian validation networks that check the presence of a shot on goal. Conditional probabilities are updated every 2 secs.

The system has been tested on MPEG-1 and MPEG-2 videos recorded at 25 frames per second (PAL standard) and with a resolution of $360 \times 288$ and $720 \times 576$, respectively. 268 case examples ($\sim$ 90 min) collected from World Championship 2002 and European Championship 2004 have been used to test the annotation system; the test set was composed by:

- 172 highlights that have been concluded with a shot on goal (SOG): 134 attack actions (AA) and 38 Placed kicks (PK)
- 54 highlights that have not been concluded with a shot on goal (NSOG): 51 attack actions and 3 Placed kicks
- 42 Other Actions (OA)

Figures of precision and recall that have been measured over the test set are reported in Table 1.

| Highlight | Precision | Recall |
|---|---|---|
| AA | 0.98 | 0.88 |
| PK | 0.63 | 0.91 |
| SOG | 0.96 | 0.88 |
| NSOG | 0.74 | 0.80 |
| OA | 0.77 | 0.95 |

**Table 1**. Performance figures of the highlight annotation engine for Attack Action (AA), Placed Kick (PK), Shot on Goal (SOG), Not Shot on Goal (NSOG) and Other action (OA)

## 3. PICTORIALLY ENRICHED ONTOLOGIES

The linguistic ontology (see Fig. 1) 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. Highlights, players and playground objects that are recognized by the annotation engine are associated with the concepts of the linguistic ontology.

In order to distinguish the specific patterns of the principal highlights detected by the annotation engine we use 6 additional visual features that are not *per se* useful for highlight classification but have instead enough discriminatory power to distinguish highlight sub-classes:

- 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.

In more detail, the playfield area is divided in twelve zones, using playfield lines and shape (see [4]). The estimation of the number of players in the upper and lower portion of the playfield (according to the playfield area that is framed) is obtained by applying a template matching of players' blobs; motion intensity and direction are extracted as described in Sect. 2; camera acceleration is computed from motion data. For each clip we create a feature vector $V$ of 6 distinct components, each of which is a vector $U$ that contains the changes within the clip of one 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.

Prototypes of the highlight patterns are obtained by clustering vectors $V$ and taking the centers of the clusters as representatives of the patterns. They are regarded as visual concepts that visually represent the specific development pattern of the highlight. Pictorially enriched ontologies are hence created by adding the prototype clip as a specialization of the linguistic concept that describes the highlight. Visual concepts in the pictorially enriched ontology are *abstractions* of video elements and can be of different types:

- *Seqs*, (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 ).

Different visual concepts can be added, incrementally, as specializations of each highlight class so as to account for the visual diversity of the highlight patterns. As a new clip is presented to the annotation system, the clustering process determines whether it belongs to existing clusters or if a new cluster must be generated. We have employed the fuzzy c-means (FCM) clustering algorithm, [12], 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 clip instances has been computed according to the Levenshtein edit distance 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 feature values. The clustering process generates the pictorially enriched ontology providing the creation of subclasses for each highlight and the creation of new highlight classes that were not defined in the initial linguistic ontology as well as the visual concepts related to each class and subclass including the visual features in the ontology. At the same time annotation of clips up to the pattern specification level is achieved by grouping clips in highlight subclasses that represent a specific visual concept.

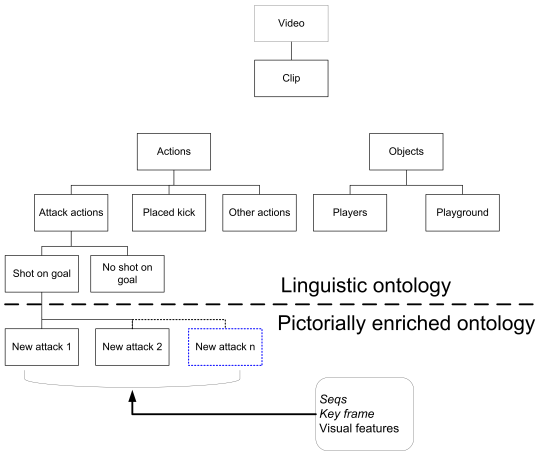


**Fig. 1**. Pictorially enriched ontology

## 4. EXPERIMENTAL RESULTS

We have performed experiments of automatic generation of pictorially enriched ontologies from video clips of soccer highlights that are automatically annotated. We have employed 40 video sequences taken from the latest Soccer World Championship. Each sequence contains number of clips variable from 3 to 8, for a total number of 258 clips. Each clip has been automatically annotated. We have focused on attack action highlights that can be terminated with a shot on goal, in that they present the largest variability of highlight patterns. Each time that a new clip is analyzed, according to the fuzzy C mean clustering algorithm, the system checks whether to assign it to an existing visual concept (the center of the clusters already detected) or if a new visual concept has to be added as a new subclass of the attack action highlight (a cluster splitting is needed). The generation of visual concepts that represent prototypes of highlight patterns has been analyzed by comparing the results obtained with the manual classification of the same highlight patterns by three human testers. Precision and recall for each cluster are reported in Table 2 where clips are considered as non relevant if they have not been assigned to the cluster by human testers. Differences in the clustering between the system and human testers are in that cluster 6 contains 2 clips that should have been split instead into two classes of 1 clip each. Similarly, cluster 8 has only 1 clip instead of 2.

Average values of precision and recall calculated over all the clusters are 0.85 and 0.83, respectively. Fig. 2 shows an example of clip clustering. We have put into evidence the clip that has been chosen as the visual concept of the highlight pattern represented by the cluster (cluster 2 of table 2), two other clips of the cluster and

one clip (enclosed in the rounded rectangle) that should have been associated with the cluster but was instead assigned to a different cluster.

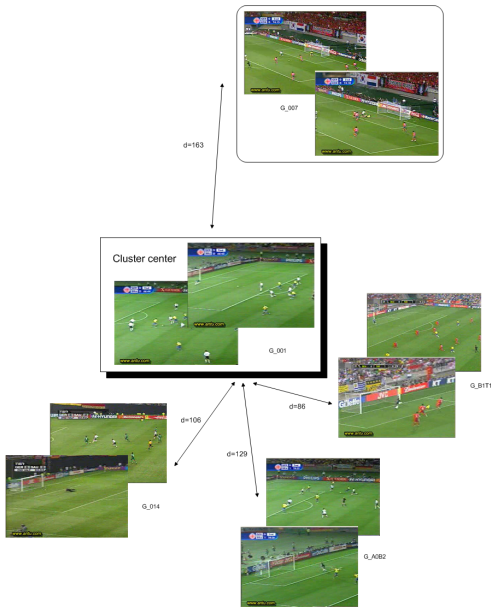


**Fig. 2**. Cluster 2 with its prototype clip (the cluster center), three clips of the cluster and one clip (enclosed in the rounded rectangle) that has been associated with a different cluster although it should belong to cluster 2. Distances w.r.t. the cluster center are indicated

## 5. CONCLUSIONS

This paper presents pictorially enriched ontologies as an extension of linguistic domain ontologies with visual features and provides a solution for their implementation in soccer video domain. A clustering algorithm has been proposed in order to create new subclasses of highlights representing specific patterns of the events and to group the clips within highlights subclasses according to their visual features. Results for automatic generation of pictorially enriched ontologies have been presented in terms of precision and recall for each highlights subclasses generated by our prototype. Experiments have shown that with pictorially enriched ontologies it is possible to extend the initial knowledge of the domain, adding subclasses of highlights or new highlight classes that were not defined in the linguistic ontology, and support automatic clips annotation up to the level of detail of pattern specification. Directions for future works are improving visual features and metrics for clustering and introducing reasoning for subclass creation and ontology enrichment.

| Cluster | Elements | Relevant | Non rel. | Precision | Recall |
|---|---|---|---|---|---|
| 1 | 6 (15%) | 5 | 1 | 0.83 | 0.83 |
| 2 | 4 (10%) | 4 | 0 | 1 | 0.8 |
| 3 | 6 (15%) | 5 | 1 | 0.83 | 0.83 |
| 4 | 11 (28%) | 9 | 2 | 0.82 | 0.9 |
| 5 | 4 (10%) | 4 | 0 | 1 | 0.8 |
| 6 | 2 (5%) | 1 | 1 | 0.5 | 1 |
| 7 | 6 (15%) | 5 | 1 | 0.83 | 1 |
| 8 | 1 (3%) | 1 | 0 | 1 | 0.5 |

**Table 2**. Precision and recall of clip clustering


### 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] World Wide Web Consortium, "Resource description framework (rdf)," Tech. Rep., W3C, http://www.w3.org/RDF/, Feb 2004.

[2] R. Leonardi and P. Migliorati, "Semantic indexing of multimedia documents," *IEEE Multimedia*, vol. 9, no. 2, pp. 44–51, April-June 2002.

[3] A. Ekin, A. Murat Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796–807, July 2003.

[4] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," *Computer Vision and Image Understanding*, vol. 92, no. 2-3, pp. 285–305, November-December 2003.

[5] X.Yu, C. Xu, H.W. 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*, Berkeley, CA (USA), 4-6 Nov. 2003 2003, vol. 3, pp. 11–20.

[6] 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.

[7] V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, and M.G. Strintzis, "Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 606–621, 2004.

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

[9] A. Jaimes and J.R. Smith, "Semi-automatic, data-driven construction of multimedia ontologies," in *ICME*, 2003.

[10] and Shih-Fu Chang Ana B. Benitez, "Automatic multimedia knowledge discovery, summarization and evaluation," *IEEE Transactions on Multimedia, Submitted*, 2003.

[11] S. Handschuh-S. Staab S. Simou N. Tzouvaras V. Petridis K. Kompatsiaris I. Strintzis, JM.G. Bloehdorn 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] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.