Dynamic Pictorially Enriched Ontologies for Video Digital Libraries

Marco Bertini[†], Rita Cucchiara[‡], Alberto Del Bimbo[†], Costantino Grana[‡], Giuseppe Serra[†], Carlo Torniai[†], Roberto Vezzani[‡]
[†]Università di Firenze - [‡]Università di Modena e Reggio Emilia

Abstract

Ontologies are traditionally used to express a conceptual view of the application domain. But for visual data, having ontologies that only provide a conceptual view is substantially inadequate for the complete expression of the semantics embedded in the perceptual patterns and therefore for supporting rich content annotation.

In this paper, we present a framework based on the Dynamic Pictorially Enriched Ontology model for semantic annotation of video streams, where the ontology includes both concepts expressed using linguistic terms and visual prototypes, that are representatives of sets of shots containing similar visual patterns. Mechanisms for instance clustering and management of cluster temporal evolution permit the creation and updating of the visual prototypes in the ontology as new knowledge is added. The OWL formalism is used to model both concepts and visual prototypes and SWRL rules are used to perform reasoning over the ontology and improve the quality of the annotation.

I. Introduction

RETRIEVAL by content from video digital libraries requires annotation of media content at both syntactic and semantic levels. But, as multimedia archives become increasingly large – as in broadcasting industry where archives are in the range of millions of hours of broadcast programs – there is the need of formal and structured annotation and retrieval systems which allow to perform complex perceptual and semantic searches, for example to reuse excerpts for the production of new videos, with acceptable computational complexity.

Keyword-based tagging systems such as the popular Flickr or YouTube services, have proved to be simple to use but their effectiveness for retrieval is weakened by the lack of a common vocabulary and tags relationships. On the other hand, Semantic Web technologies promise to ease the task of creation, interchange and management of metadata, and support data access at semantic level. According to Semantic Web paradigm, ontologies are regarded as the formal tool to express the concepts and their attributes, as well as the relationships between concepts, in the domain of interest.

Indeed, ontologies play a fundamental role also for the efficient annotation of visual content in video digital libraries and therefore for content-based retrieval [1], insofar as they allow association between concepts and visual data [2]. Ontologies useful for semantic annotation of visual data are those defined by the Dublin Core Metadata Initiative [3], TV Anytime [4] - they have defined standardized metadata vocabularies - and the LSCOM initiative [5] - that has created a specialized vocabulary for news video. In these cases, ontologies include a set of linguistic terms with their associated definitions that formally describe the application domain, through concepts, concept properties and relations, according to some particular view.

However, linking ontology concepts to visual data poses a number of problems that are still far to be solved despite of the research efforts spent so far. A first problem is concerned on how to obtain a complete expression of the information content of visual data. In many cases, such as for complex scenes or events, the usage of linguistic concepts alone is substantially inadequate for the complete expression of the semantics embedded in visual data. According to Gardner's studies in cognitive psychology, the basis of the cognition process is the "representation", describing the different modalities of mental representations such as symbols, images or schemata [6]. Therefore, both perceptual (based on low-level features) patterns

and semantic concepts are necessary. Another important problem is related to the fact that objects and events can change their visual manifestations through time. This reflects into the fact that the association between visual data and high-level concepts should have some temporal evolution that directly reflects on the interpretation of visual content.

In this paper we present a framework for annotation of video streams based on an ontology model, referred to as Dynamic Pictorially Enriched Ontology that addresses these problems. In this model, in addition to linguistic concepts, visual prototypes are included as representatives of visual patterns accounting for the different modalities in which visual data can manifest. They are obtained by clustering the instances of visual data that are observed, according to distinguishing perceptual features. The temporal modification of visual data is managed through a clustering mechanism that, when a new data instance is presented to the ontology, determines re-clustering of the instances already observed and re-definition of the visual prototypes. The Web Ontology language (OWL) formalism is used to model both domain concepts and visual prototypes, and rules expressed with Semantic Web Rule Language (SWRL) are used to perform reasoning over the ontology so as to enhance the results of the classification or derive new semantic annotations of shots.

II. Previous Work

Some previous work on ontologies has focused on the use of linguistic ontologies and appropriate classifiers to perform concept association to visual data. In these approaches the ontology provides the conceptual view of the domain at the schema level, and the classifiers play the role of observers of the real world sources. They classify the observed entity in the nearest concept of the ontology, implementing invariance with respect to several conditions. Once the observed entity is classified, the ontology is exploited to have a richer semantic annotation, establishing links to the high-level concepts and disambiguating the result of classification. Snoek et al. [7] proposed a method to perform video annotation with the MediaMill 101 concept lexicon. In this work machine learning technique trains classifiers to detect high-level concepts from low-level features, while WordNet is used to derive high-level concepts relations in order to enhance the annotation performance. Zha et al. [8] have defined an ontology to provide some structure to the LSCOM-lite lexicon, using pairwise correlations between concepts and hierarchical relationships, to refine concept detection of SVM classifiers.

Other authors have postulated the idea of including visual data instances in the ontology so as to account for the variety of manifestations of visual information. In this approach, feature detectors are applied to raw data of the external source and the features extracted are matched against those of the concept instances in the ontology, in order to link the observations with the concepts of the schema. Petridis et al. [9], defined a Visual Descriptors ontology, a Multimedia Structure ontology and a Domain ontology to perform video content annotation at semantic level. The Visual Descriptors ontology included concept instances represented with MPEG-7 visual descriptors. Kompatsiaris et al. [10], included in the ontology instances of visual objects. They used as descriptors qualitative attributes of perceptual properties like color homogeneity, low-level perceptual features like components distribution, and spatial relations. Semantic concepts were derived from color clustering and reasoning. Maillot and Thonnat [11] have proposed a visual concept ontology that includes texture, color and spatial concepts and relations for object categorization. Dasiopoulou et al. [12] proposed an ontology-based framework for enhancing segment-level annotations resulting from typical image analysis, through the exploitation of visual context and spatial relations.

In the attempt of having richer annotations, other authors have explored the usage of reasoning over multimedia ontologies. In this case spatio-temporal relationships between concept occurrences are analyzed so as to distinguish between scenes and events and provide more fitting and comprehensive descriptions. Similar approaches have been used in Espinosa et al. [13] and Neumann and Möller [14]. In the former, inference from observation to explanation (abduction) is used to check, among detected entities, relations and constraints that lead to consistent interpretation of image content; the latter exploits reasoning over aggregates, defined as sets of objects to be recognized as a whole, for scene interpretation.

The inclusion of data instances in the ontology requires some mechanism for the management of the ontology evolution. In the EC project Boemie the authors have addressed the problem of temporal evolution of visual data [15]. Each visual instance is checked in order to determine whether it can be associated to the existing abstract concepts or a new concept has to be defined in the ontology. Evolution patterns have been proposed to define the kinds of action to be performed over the ontology: instance population, leveraging detected mid and high-level concepts relations to perform annotation or ontology evolution, defining new concepts and enriching the domain ontology with these new concepts and their relations.

Differently from these approaches our proposed Dynamic Pictorially Enriched Ontology framework addresses the issues stemming from integrating visual information in the ontology by i) including visual instances related to high-level concepts and identifying their spatio-temporal patterns, ii) defining visual prototypes representative of these patterns using them for automatic annotation, and iii) supporting visual prototypes evolution in time. Moreover, our approach emphasizes the need of spatio-temporal constraints among objects and entities for complex video content interpretation and proposes the use of SWRL as an effective mean to define, share and refine rules that can lead to concepts definition and recognition within video content.

III. Automatic Video Annotation Framework

Our framework implements methods for the extension of linguistic ontologies with visual information, employs the Dynamic Pictorially Enriched Ontology model to perform video annotation, and leverages cluster updating to support temporal evolution of the visual prototypes. In this model the ontology contains linguistic domain concepts, their relationships and visual instances. Concepts that are selected to have visual instances are those that have changes in shape, appearance and motion in their spatio-temporal pattern. In the ontologies used in the experiments these concepts comprise sport highlights such as shot on goal and pitstop, or views such as long-range and close-up views. Visual instances, associated to the concepts of the schema, include object identifier, visual descriptors, time label and link to the raw data (e.g. shot). These instances are created as the result of the matching between the descriptors of the raw visual data and the descriptors of one reference instance in the ontology. Clustering is used to group instances that have some similarity in their visual or spatio-temporal patterns; in general several clusters exist for each concept, each of which roughly corresponds to one modality in which that concept can manifest itself in the reality.

In order to reduce the cost of descriptors matching, for each cluster, a visual prototype is defined that is assumed to represent all the instances in the cluster. The median element is assumed as the visual prototype of the cluster. Visual prototypes are created initially using a training set of already annotated data, performing clustering according to their visual descriptors.

A special cluster, the *unknown concept* cluster, is created that includes all the instances that have not been assigned yet to a cluster. In consideration of the fact that the visual instances may have a large variety of modes in which they appear, any new instance that is presented to the ontology for annotation can be regarded as new knowledge for the ontology. According to this, every time a new instance is associated to a concept, cluster updating is performed over all the clusters of that concept and the unknown concept cluster. In this way, previously labeled unknown instances can be assigned to some cluster, or new clusters are created as a result of the new instance. Ultimately this permits to represent more effectively the variety of appearances and motion patterns of the instances and provides a form of temporal evolution for the knowledge in the ontology. An example that shows some of the possible clustering updates is schematized in Fig. 1.

Visual instances are obtained by video segmentation and feature extraction. Video segmentation is performed by a shot detection algorithm called LTD (Linear Transition Detection); this technique approximates a linear transition model and is quite robust to detect cuts, dissolves and fades [16]. The descriptors extracted from video shots are of two types: *low-level features*: i.e. visual descriptors such as color histograms, edge maps, etc., that are usually related to concepts that may be used in different

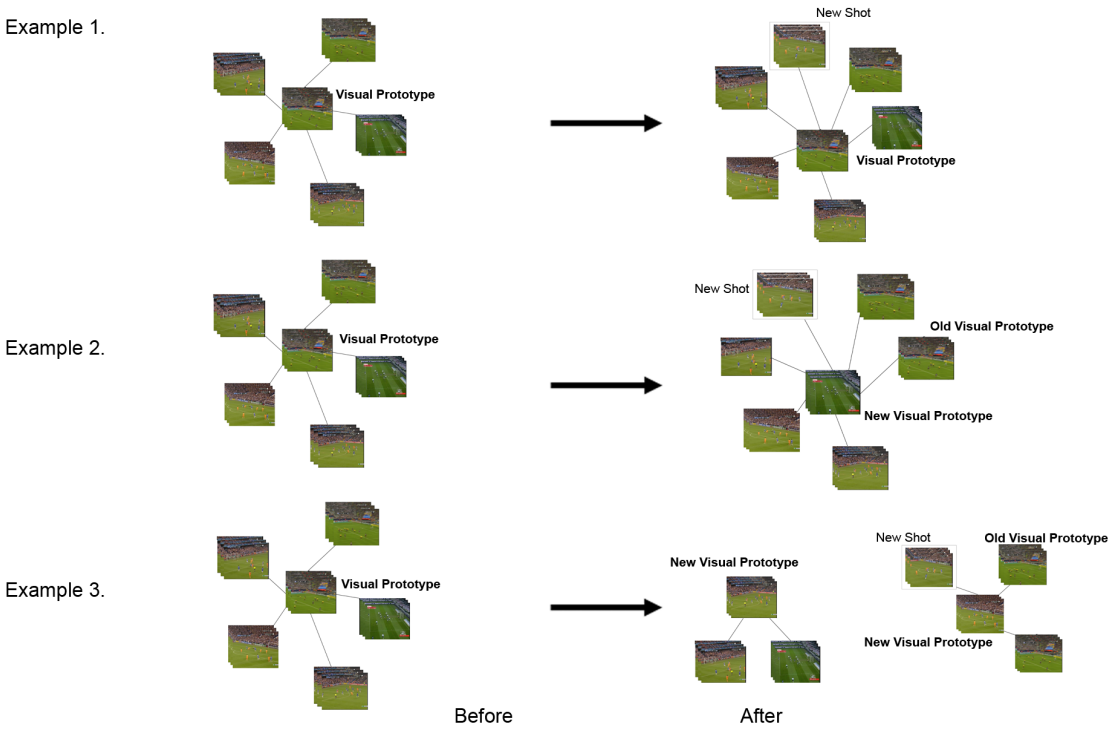


Fig. 1: When a new shot is assigned to a concept all the clusters of the concept are updated; some possible outcomes of the updating are shown: example 1) the visual prototype of the cluster remains the same, example 2) a new prototype is selected, example 3) the cluster is split and two new visual prototypes are selected.

domains (such as scene setting, shot type, indoor/outdoor context, etc.); *medium-level features*: i.e. visual descriptors that refer to the particular application domain (such as playfield position and area in sport videos, anchorman frames in news videos, etc.), or descriptors of specific entities (such as faces obtained from face detection, text extracted from superimposed-text detector, etc.). We used literals to represent the sequences of the values of descriptors (computed from each frame of the shot) to account for different lengths of the shots and for the temporal order of the values.

To experiment its capability in semantic annotation of video streams, the proposed framework was applied to the Formula 1 and soccer video domains. Fig. 2 shows part of the schema of the soccer ontology, with linguistic concepts and visual instances. For each cluster of visual instances, the *has_visual_instance* property allows linking to visual instances; the *has_visual_prototype* property instead identifies the cluster visual prototype. The annotation process (matching of observed data instances with visual prototypes and consequent high-level concept association) and cluster updating process (the observed data instances determines re-clustering of the clusters in the ontology) are also schematized.

Different clustering methods have been used. For concepts where temporal ordering of their descriptors is not relevant (typically they refer to scene views such as close-up or wide angle views) we used the Complete Link hierarchical clustering. The Mallow's distance $d(S_x, S_y)$ was used as a measure of the dissimilarity between the features vector of two shots S, so as to account for shots of different length. The process begins assigning every shot to a different cluster, then proceeds with an iterative grouping of the two most similar clusters until a single cluster that contains all the instances is obtained. Each level of the hierarchy can be identified by its number n of clusters. The optimal number $\tilde{n}$ of clusters is automatically obtained by maximizing a Clustering Score CS_n. To this aim we define the diameter $\Delta(W_i)$ of a cluster

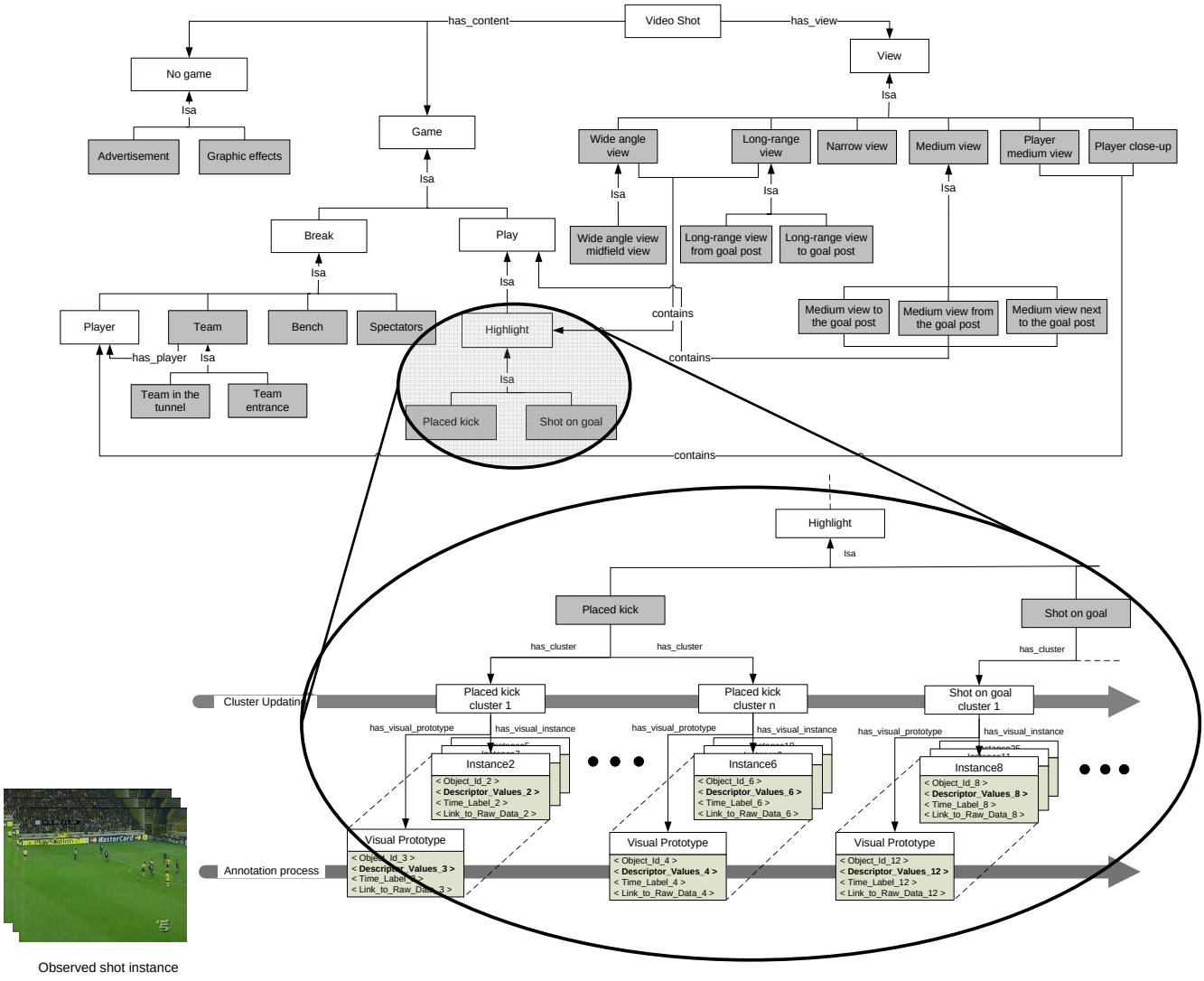


Fig. 2: The video annotation framework with Dynamic Pictorially Enriched Ontologies showing: partial schema of the soccer video ontology with linguistic concepts and visual instances; instance clusters with visual prototypes; annotation by matching the observed instance with visual prototypes and cluster updating.

W_i and the distance $\delta(W_i, W_j)$ between two clusters W_i and W_j as:

$$\Delta(W_i) = \max_{S_x, S_y \in W_i} d(S_x, S_y)$$

$$\delta(W_i, W_j) = \min_{S_x \in W_i, S_y \in W_j} d(S_x, S_y)$$

The Clustering Score at level n is thus defined as:

$$CS_n = \min(\Delta_1 - \Delta_n, \delta_n)$$

where

$$\Delta_n = \max_{W_i \in E_n} \Delta(W_i)$$

$$\delta_n = \min_{W_i, W_j \in E_n, i \neq j} \delta(W_i, W_j)$$

being E_n the set of clusters at level n.

On the other hand, for concepts where the temporal evolution of descriptors is extremely relevant (for example in a soccer shot on goal action, where the order of players motion and zone of playfield values are important elements to distinguish one action from the other) we used Fuzzy C-Means (FCM) clustering method and the Needleman-Wunch distance. This distance accounts for the fact that shots may have different temporal length and accounts for the temporal order of the features. It is obtained as the sum of all the normalized Needleman-Wunch distances between the distinct components of the content descriptors, is defined as follows:

$$d(S_x, S_y) = \frac{\sum_U NW(U_{S_x}, U_{S_y})}{min(length(S_x), length(S_y))}$$

where U is a vector obtained as the composition of the individual content descriptors of the shot [17].

Some concepts that may occur in a video can not be detected and recognized only from the observation of visual features. In some cases, instead, they can be recognized analyzing the context and the content of the preceding and following shots. For example, in the soccer domain some placed kick that are not recognized using the visual features, can be recognized considering that they are often preceded by player close-up and medium view shots (showing the player that is going to kick the ball, and the other players). We can therefore define patterns that use temporal relations between concepts to improve shot annotation. In our framework we have used rules to model these patterns and rule-based reasoning for their recognition. Patterns can include conditions on the occurrence of concepts, constraints on the values of visual descriptors of concept instances and temporal relations between concepts occurrences. SWRL is used to define the rules that model these patterns. With respect to OWL axioms, SWRL permits to express rules explicitly through *if-then* expressions, and provides built-in mathematical, logical, string comparison and temporal operators, making it easier the definition and modification of rules and rule constraints, even by non-programmer domain experts [18].

IV. Experimental results

We have experimented our annotation framework on Soccer and Formula 1 domains, so as to show its general applicability and analyse the performance improvements achievable. Table I lists the concepts with visual prototypes that have been selected in the two domains. Typically they represent dynamic actions or highlights (such as shot on goal or pitstop exit), views commonly used for the overview of an event (such as long-range and wide-angle views) and views of the surrounding context (such as team and spectators). For the description of the visual content, we used the same set of low-level MPEG-7 visual features for the concepts of the two domains, namely: Color Layout, Scalable Color, Edge Histogram and Motion Activity descriptors. For the Soccer domain highlights we added a few other medium-level features, namely: the main camera motion direction and intensity (that approximately model the motion of the ball), the playfield zone framed, and the number of visible players in the upper and lower part of the playfield. The shot on goal and placed kick highlights were described explicitly considering the temporal evolution of their visual features. In the following we show the results of the experiments with Dynamic Pictorially Enriched Ontologies.

The first experiment checks performance of video annotation in dependence on the number of the shot instances in the training set that is used to create the visual prototypes. The analysis was made on soccer videos for two sets of concepts, those that don't exploit any temporal information and those that consider the temporal order of the visual features observed. For the first set of concepts, we have used six distinct collections of video sequences extracted from World-championship 2006. They contain different games and athletes, are taken in different stadiums, have different lengths and many different edit effects. Each collection includes a few concepts such as wide angle view, long-range view, medium view, spectators, team, player close-up, bench or their specializations. Results of the experiment for this set are reported

TABLE I: List of concepts with visual prototypes, used in the soccer and Formula 1 video domain ontologies. The indentation indicates that a concept is a specialization of another concept.

Soccer domain	**Description**
Wide angle view	wide angle view framed by the main camera
Wide angle midfield view	wide angle view of the midfield area
Long-range view	long-range view focussing over middle area
Long-range view from goal post	long-range view as taken from the goal posts
Long-range view to the goal post	long-range view taking the goal post in the center
Narrow view	narrow angle view as taken from some handheld video camera
Medium view	players are fully displayed in the playfield
Medium view from the goal post	medium view as taken from the goal post
Medium view to the goal post	medium view as taken from the goal post
Medium view next to the goal post	medium view where the goal post is lateral in the image
Players medium view	players view framed from a side camera near the playfield
Bench	coach and team staff view
Player close-up	players close-up view
Team	view of the team
Team entrance	view of the team entrance in the play field
Team in the tunnel	view of players taken in the tunnel before and after the game
Spectators	view displaying supporters and cheering crowd
Advertisement	view showing advertisement
Graphic effects	view displaying any other elements such as computer graphics
Shot on goal	action where a player kicks the ball to the opponent's goal post to score a goal
Placed kick	action including penalty, corner and free kick near the goal post

Formula 1 domain	**Description**
Wide angle view	wide angle view of the race track
Medium view	medium view of the track
Car close-up	car close-up view
Box staff	staff view
Spectators	view displaying supporters and cheering crowd
Advertisement	view showing advertisement
Camera-car driver view	view of the driver from the camera car
Camera-car front view	view of the front of the car from the camera car
Box pitstop	view displaying a car and the team during the pitstop
Box car entry	action where a car is entering the box
Box car exit	action where a car is exiting from the box
Race start	action at the start of the race

in Table II(a) in terms of average precision and recall. Initially (row 1), 81 shots were presented to the initial ontology (trained with 25 shots annotated manually to create the initial set of visual prototypes representing 7 concepts). The 81 shots were annotated exploiting the initial set of visual prototypes and manually for the concepts that were not already represented (11 concepts represented in total). This process was repeated in several steps, with the addition of new shots at every step. As new shots are presented to the system, the visual prototypes in the ontology are updated consequently. At step 5, 1158 shots were used containing 19 concepts. The improvement of the quality of annotation can be observed considering the figures of precision and recall from row 2 to 5. The second set of concepts includes shot on goal and placed kick. These concepts have a large variability of their visual appearance, so that a large number of visual prototypes is necessary to permit effective annotation. According to this, initially (row 1) 68 shots were presented to the system ontology trained with 30 shot on goal and 20 placed kick shots (row 1). In each of the following steps (row 2 and 3) 50 additional shots were used for training. Reasonably, the experiment shows that increasing the training set for the creation of the visual prototypes of the ontology concepts results into improvement of performance; in particular, the improvement of recall is mainly due to the fact that the number of shots classified as unknown concept decreases as the number of visual prototypes increases.

The second experiment highlights the capability of Dynamic Pictorially Enriched Ontologies to capture the temporal evolution of the visual prototypes. To this end we used the already annotated placed kick

TABLE II: Evaluation of the annotation performance depending on the number of the shot instances composing the training set used for visual prototype creation

(a) Annotation of concepts that do not require temporal ordering of the features.

Video collection	No. shots	No. of concepts	Shots used for ontology training	Avg. precision	Avg. recall
1	81	11	25	0.19	0.20
2	206	17	106	0.34	0.28
3	255	19	312	0.47	0.38
4	591	19	567	0.43	0.46
5	341	19	1158	0.55	0.52

(b) Annotation of concepts that require temporal ordering of the features.

Video collection	No. shots	No. of concepts	Shots used for ontology training	Avg. precision	Avg. recall
1	68	2	50	0.43	0.27
2	68	2	100	0.53	0.40
3	68	2	150	0.62	0.60

TABLE III: Dynamic evolution of visual prototypes in the ontology for the domain specific placed kick highlight (2001-2006)

Placed kick action					
Video collection	Years	Mean shift	σ^2 shift	Mean cluster radius	n. visual prototypes
1	2001			3.6	4
2	2001+2005	3.3	1.3	4.5	6
3	2001+2005+2006	1.9	1.7	4.4	8

shots used for training in the previous experiment. Since they collected events filmed in years 2001, 2005 and 2006, they were input to the ontology in three distinct steps, so that clusters are updated at each step with re-definition of the visual prototypes at the cluster centers. While feeding the ontology with these data we also kept track of the way in which the visual prototypes of placed kicks changed in the ontology. In particular, at each step, we registered the mean and variance of the shifts of the cluster centers with respect to their position at the previous step, together with the mean radius of each cluster and the number of clusters. In the case of cluster splitting, the shift of the original cluster center is calculated with respect to the closest of the new cluster centers. Table III (row 1,2 and 3) shows the evolution of these parameters for each step. Distances were calculated according to the Needleman-Wunch distance defined in Sect. III. Temporal evolution of cluster prototypes implicitly provides additional semantic information about the way in which modifications of phenomena occur. For the case of example it provides some evidence of the different modes in which placed kicks have been filmed and displayed in TV from 2001 to 2006. Actually, placed kick shots of 2005 determine a large average shift of the 2001 placed kick cluster centers, with an increase of both the mean cluster radius and the number of clusters (they determine big changes in the visual prototypes of 2001) (see row 2); besides, adding 2006 placed kick shots results into the increase of the number of visual prototypes, but with a smaller average shift of the cluster centers and a slight decrease of the mean cluster radius (they have more similar temporal evolution of the visual features as 2005 shots) (see row 3). Indeed, this reflects the fact that camera shooting was changed considerably starting from 2005: the long phase of preparation of the kick (placing the ball, waiting for the placement of the opponents, etc.) is now shown rarely, being replaced by players' close-ups and medium views of the playfield, so as to display a faster and more dynamic highlight scene.

In the the third experiment we measured performance of annotation with Dynamic Pictorially Enriched Ontologies with respect to the concepts defined for the domains of Soccer and Formula 1 in Table I. Tests were performed on the same shots of the first experiment. For each concept we indicated correct, miss, false and unknown classified shots, with display of the average precision and recall achieved.

TABLE IV: Performance of annotations for the concepts defined in Table I

(a) Soccer domain

Concept	Correct	Unknown	Miss	False	Precision	Recall
Wide angle view	26	0	5	14	0.65	0.84
Wide angle midfield view	1	0	2	3	0.25	0.33
Long-range view	12	4	9	9	0.57	0.57
Long-range view from the goal post	4	0	5	1	0.80	0.44
Long-range view to the goal post	1	0	2	3	0.25	0.33
Narrow view	2	1	7	5	0.29	0.22
Medium view	26	2	11	18	0.59	0.70
Medium view from the goal post	1	0	3	3	0.25	0.25
Medium view to the goal post	6	1	2	5	0.54	0.75
Medium view next to the goal post	1	2	4	1	0.50	0.20
Players medium view	1	1	4	5	0.16	0.20
Bench	7	1	14	3	0.70	0.33
Player close-up	72	5	27	20	0.78	0.73
Team	10	0	4	1	0.90	0.71
Team entrance	1	0	3	1	0.50	0.25
Team in the tunnel	1	1	1	4	0.20	0.50
Advertisement	5	0	1	0	1	0.83
Graphic effects	22	1	1	0	1	0.96
Spectators	15	0	3	12	0.55	0.83
Shot on goal	20	5	5	6	0.77	0.67
Placed kick	11	9	1	12	0.48	0.52

(b) Formula 1 domain

Concept	Correct	Unknown	Miss	False	Precision	Recall
Wide angle view	53	2	33	10	0.84	0.62
Medium view	47	4	31	46	0.50	0.60
Car close-up	76	4	18	54	0.58	0.81
Box staff	23	0	36	13	0.64	0.39
Spectators	41	0	43	13	0.76	0.49
Advertisement	96	3	4	8	0.92	0.96
Camera-car driver view	97	3	2	3	0.97	0.98
Camera-car front view	7	2	0	1	0.87	1
Box pitstop	57	0	44	29	0.66	0.56
Box car entry	78	0	21	20	0.80	0.79
Box car exit	81	3	19	54	0.60	0.79

For the Soccer domain, the results observed are reported in Tab. IV(a). It can be observed that some concepts, like narrow view and team, are poorly represented mainly due to the fact that they present very high variability of appearance. Player close-up shots share instead a typical appearance (a person is usually at the center of the frame, taking large part of it). High precision and recall are observed in that color layout and edge histogram descriptors are appropriate features for the representation of this concept. Most false detections observed are due to shots with narrow angle view and medium view concepts, showing a person in the foreground. In these cases there is the need of additional semantic information to recognize the concept properly. Shot on goal shows a good performance in terms of both precision and recall. Low recall and large number of unknown concept for placed kick are principally due to the fact that placed kick dataset includes placed kicks filmed with different modalities (see the second experiment). Table IV(b) reports results for the Formula 1 domain. Camera-car concepts show excellent precision and recall in that they can be easily distinguished from the other concepts. Box car entry and exit are very well characterized from motion feature. Differently, box staff and box pitstop can be easily confused each other in that they don't have an univocal characterization of motion.

In the fourth experiment, we have provided some evidence of the improvement of precision and recall that is achievable with rule-based ontology reasoning, even with the addition of few simple rules. The analysis has been performed for a few principal highlights of soccer and car racing. The Jess reasoning

TABLE V: Comparative performance analysis

(a) Precision and recall of annotation of domain specific highlight concepts with visual prototypes (VP); with visual prototypes and SWRL reasoning (VP+SWRL), and with reasoning only (SWRL).

Highlight	Shot on goal		Placed kick		Box car exit		Race start
	VP	VP+SWRL	VP	VP+SWRL	(VP)	VP+SWRL	SWRL
Correct	20	23	11	18	81	83	7
Unknown	5	2	9	2	3	1	-
Miss	5	5	1	1	19	19	3
False	6	6	12	12	54	54	2
Precision	0.77	0.79	0.48	0.60	0.60	0.61	0.78
Recall	0.76	0.77	0.52	0.86	0.79	0.81	0.70

(b) Comparison of precision and recall of annotation of domain specific highlight concepts with visual prototypes and SWRL reasoning (VP+SWRL) and SVM classifiers (SVM).

Highlight	Shot on goal		Placed kick	
	VP+SWRL	SVM	VP+SWRL	SVM
Correct	23	18	18	15
Unknown	2	-	2	-
Miss	5	12	1	6
False	6	5	12	9
Precision	0.79	0.78	0.60	0.63
Recall	0.77	0.60	0.86	0.71

engine was used in the experiments. SWRL rule patterns were defined for placed kick, shot on goal, box car exit and race start; the placed kick rule is defined as follows:

- IF player close-up shots occur before an unknown concept shot, with a few seconds of fixed camera, within a time interval between 40 and 50 seconds THEN the unknown concept shot is classified as a placed kick.

Fig. 3a) shows an example with the SWRL code for this pattern, and the rules for placed kick, box car exit and race start (Fig. 3b). The results of the comparison are shown in Table V(a). For Soccer, it appears that the SWRL rules have particularly improved recall. Instead for box car exit recall and precision remain almost the same. This was due to the fact that the number of shots classified as unknown concept was anyway very low. On the other hand it is shown that SWRL reasoning can be usefully employed to detect concepts that are characterized by some temporal structure, as the race start. In Table V(b) we compare the performance obtained through the use of SWRL rules and ontology reasoning to the traditionally employed SVM classification, for shot on goal and placed kick highlights. In order to have a fair comparison, SVM classifiers (with RBF kernel) have been trained over the same training set of the ontology. Video clips have been represented with the same vectors used for concept clustering in the ontology, with a fixed number of samples (5) per clip so as to have feature vectors of the same length. The improvement with Dynamic Pictorially Enriched Ontology is essentially due to the fact that, differently from SVM, SWRL rules permit to include some contextual information, namely temporal constraints, that allows to disambiguate situations beyond the value of visual features.

V. Conclusions

In this work we have presented a framework for video annotation based on Dynamic Pictorially Enriched Ontologies. The ontology model is defined in OWL, and includes both linguistic concepts and visual prototypes. In order to address issues related to linking ontology concepts to visual data, in the proposed framework we have implemented mechanisms for instance clustering, so as to create visual prototypes that are representatives of sets of shots with similar visual patterns, and cluster updating, so as to account for knowledge temporal evolution of visual specification of concepts. We have tested our approach in two different domains and the experiments showed its effectiveness in video annotation, in supporting

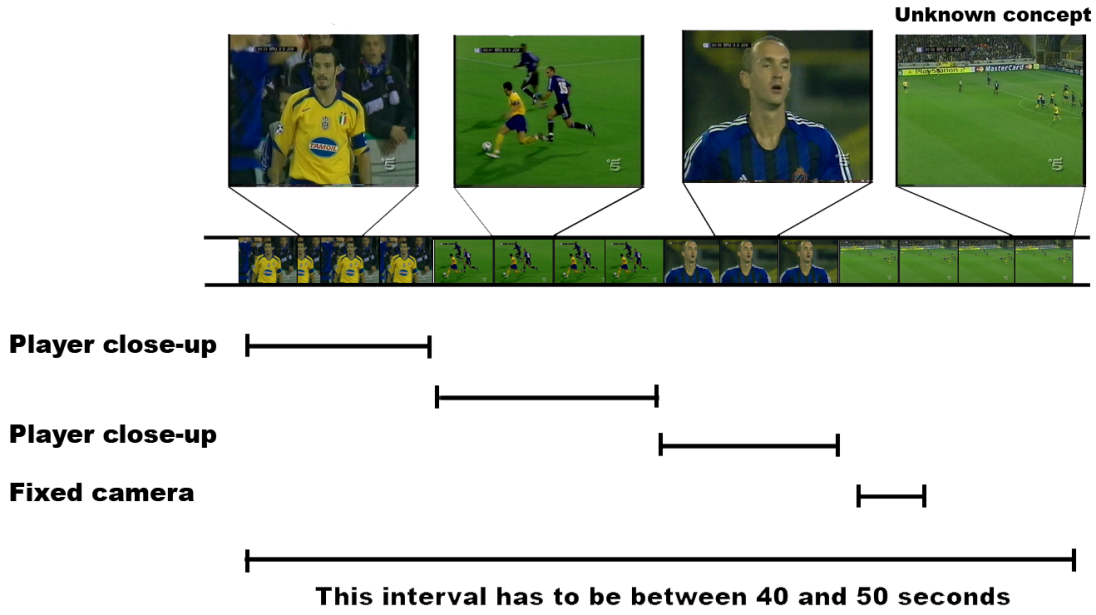
temporal evolution of concepts' visual specifications and in improving annotation performance through rule-based reasoning.

ACKNOWLEDGMENT

This work was partially supported by the IST Program of the European Commission DELOS Network of Excellence and by the VIDI-Video project.

REFERENCES

[1] J. Hunter, *Multimedia Content and the Semantic Web: Standards, methods and Tools.* John Wiley & Sons, 2005, ch. Adding Multimedia to the Semantic Web: Building and Applying an MPEG-7 Ontology.
[2] L. Hollink, G. Schreiber, and B. Wielinga, "Patterns of semantic relations to improve image content search," *Journal of Web Semantics*, vol. 5, no. 3, pp. 195–203, Sept. 2007.
[3] Dublin core metadata initiative. [Online]. Available: http://dublincore.org/
[4] TV anytime forum. [Online]. Available: http://www.tv-anytime.org/
[5] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, July-Sept. 2006.
[6] H. Gardner, *The mind's new science: a history of the cognitive revolution*. Basic Books, Inc., 1985.
[7] C. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE TMM*, vol. 9, no. 5, pp. 975–986, Aug. 2007.
[8] Z.-J. Zha, T. Mei, Z. Wang, and X.-S. Hua, "Building a comprehensive ontology to refine video concept detection," in *Proc. MIR*, Sept. 2007, pp. 227–236.
[9] S. Bloehdorn, N. Simou, V. Tzouvaras, K. Petridis, S. Handschuh, Y. Avrithis, I. Kompatsiaris, S. Staab, and M. G. Strintzis, "Knowledge representation for semantic multimedia content analysis and reasoning," in *Proc. EWIMT*, London, U.K., Nov. 2004.
[10] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, and M. G. Strintzis, "Knowledge-assisted semantic video object detection," *IEEE TCSVT*, vol. 15, no. 10, pp. 1210–1224, 2005.
[11] N. Maillot and M. Thonnat, "Ontology based complex object recognition," *Image Vision Comput.*, vol. 26, no. 1, pp. 102–113, 2008.
[12] S. Dasiopoulou, C. Saathoff, P. Mylonas, Y. Avrithis, Y. Kompatsiaris, S. Staab, and M. Strintzis, *Semantic Multimedia and Ontologies Theory and Applications*. Springer, 2008, ch. Introducing Context and Reasoning in Visual Content Analysis: An Ontology-Based Framework.
[13] S. Espinosa, A. Kaya, S. Melzer, R. Moller, and M. Wessel, "Towards a media interpretation framework for the semantic web," in *Proc. ICWI*, 2007, pp. 374–380.
[14] B. Neumann and R. Moeller, "On scene interpretation with description logics," in *Cognitive Vision Systems: Sampling the Spectrum of Approaches*, ser. LNCS. Springer, 2006, pp. 247–278.
[15] S. Castano, S. Espinosa, A. Ferrara, V. Karkaletsis, A. Kaya, S. Melzer, R. Moller, S. Montanelli, and G. Petasis, "Ontology dynamics with multimedia information: The BOEMIE evolution methodology," in *Proc. IWOD*, 2006.
[16] C. Grana and R. Cucchiara, "Linear transition detection as a unified shot detection approach," *IEEE TCSVT*, vol. 17, no. 4, pp. 483–489, 2007.
[17] M. Bertini, R. Cucchiara, A. Del Bimbo, and C. Torniai, "Video annotation with pictorially enriched ontologies," in *Proc. ICME*, 2005.
[18] L. Hollink, S. Little, and J. Hunter, "Evaluating the application of semantic inferencing rules to image annotation," in *Proc. K-CAP*, 2005.



`Unknown_concept(?u)` ^	instance of unknown action
`has_Valid_Period(?u,?Vpu)` ^	interval time of the unknown action
`temporal:hasFinishTime(?Vpu,?Ftu)` ^	end time of the unknown action
`Player_close_up(?p1)` ^	instance of close-up *p1*
`Player_close_up(?p2)` ^	instance of close-up *p2*
`differentFrom(?p1, ?p2)` ^	the two close-ups have to be different
`has_Valid_Period(?p1, ?Vpp1)` ^	interval time of the close-up *p1*
`temporal:hasStartTime(?Vpp1,?Stp1)` ^	start time of the close-up *p1*
`has_Valid_Period(?p2, ?Vpp2)` ^	interval time of the close-up *p2*
`temporal:before(?Vpp1,?Vpp2)` ^	close-up *p1* has to start before the close-up *p2*
`Fixed_Camera(?f)` ^	instance of fixed camera
`has_Valid_Period(?f, ?Vpf)` ^	interval time of the fixed camera
`temporal:before(?Vpp1, ?Vpu)` ^	interval time of the close-up *p1* has to be before the interval time of unknown action
`temporal:before(?Vpp2, ?Vpu)` ^	interval time of the close-up *p2* has to be before the interval time of unknown action
`temporal:starts(?Vpu,?Vpf,temporal:Seconds)` ^	interval time of unknown action and fixed camera have to start together
`temporal:duration(?diff, ?Stp1, ?Ftu, temporal:Seconds)` ^	duration between start time of close-up *p1* and the end time of unknown action
`swrlb:greaterThan(?diff, 40)` ^	the duration has to be greater than 40 seconds
`swrlb:lessThan(?diff, 50)` ^	the duration has to be less than 50 second
->	
`Placed_Kick(?u)`	then the instance of unknown action is a placed kick

(a) Placed kick SWRL rule.

Shot on goal: *IF player close-up shots AND medium view to the goal post occur after an unknown concept shot, with a few seconds of goal post view, within a time interval between 10 and 20 seconds THEN the unknown concept shot is classified as shot on goal.*

Box car exit: *IF box pitstop OR box staff shots occur before an unknown concept shot, with a few seconds of motion activity AND wide angle OR medium view follow within a time interval between 7 and 20 seconds THEN the unknown concept shot is classified as box car exit.*

Race start: *IF camera-car front view AND car close-up occur before medium view, without motion activity, within a time interval between 50 and 70 seconds THEN the medium view shot is classified as race start.*

(b) Shot on goal, box car exit and race start rules.

Fig. 3: a) Example of placed kick pattern. A placed kick shot is detected if: player close-up shot(s) are followed by an unknown concept shot, with few seconds of fixed camera, within a time interval from 40 to 50 seconds. b) Rules for shot on goal, box car exit and race start.