Dynamic Pictorially Enriched Ontologies for Digital Video Libraries

Marco Bertini, Alberto Del Bimbo, Giuseppe Serra, and Carlo Torniai
University of Florence, Italy

Rita Cucchiara, Costantino Grana, and Roberto Vezzani
University of Modena and Reggio Emilia

This article presents a framework for automatic semantic annotation of video streams with an ontology that includes concepts expressed using linguistic terms and visual data.

Retrieval by content from video digital libraries requires annotation of media content at both syntactic and semantic levels. As multimedia archives become increasingly large—some broadcasting archives contain millions of hours of footage—the need for more sophisticated annotation and retrieval systems becomes increasingly acute. Keyword-based tagging systems, such as those used by Flickr or YouTube, are simple to use but lack a common vocabulary and tag relationships, reducing their retrieval effectiveness. Semantic Web technologies promise to make managing the metadata associated with these archives much easier. In the Semantic Web paradigm, ontologies become the formal tool to express concepts, their attributes, and the relationships between concepts in the domain of interest.

Indeed, ontologies can play a fundamental role in efficiently annotating content in digital video libraries because they allow association between concepts and visual data.[1,2] For semantic annotation of visual data, there are several existing and useful ontologies, including those defined by the Dublin Core Metadata Initiative (see http://dublincore.org), by TV Anytime (see http://www.tv-anytime.org), and by the Large Scale Concept Ontology for Multimedia initiative.[3] In these cases, the ontologies include a set of linguistic terms and definitions that formally describe the application domain through concepts, concept properties, and relations, all according to some particular view.

However, linking ontology concepts to visual data poses several problems that are still far from being solved. One key problem is how to obtain a complete expression of the information content of visual data. In many cases, such as for complex scenes or events, using linguistic concepts alone is inadequate for a complete expression of the semantics embedded in visual data. According to studies in cognitive psychology, the basis of the cognition process is the different modalities of mental representations, such as symbols, images, and schemata.[4] Therefore, both perceptual patterns and semantic concepts are necessary for a complete expression of visual data. Another key problem is the fact that the visual manifestations of objects and events can change over time, which suggests that the association between visual data and high-level concepts should have some kind of built-in temporal evolution.

In this article, we present a framework for annotating video streams with an ontology model designed to address these problems. The Dynamic Pictorially Enriched Ontology model includes not only linguistic concepts, but also visual prototypes to account for visual data's 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 model manages temporal modification of visual data through a clustering mechanism that can recluster instances already observed and redefine the visual prototypes. We use the Web Ontology Language to model both domain concepts and visual prototypes, and the Semantic Web Rule Language (SWRL) to enhance, through reasoning, the results of the classification and derive new semantic annotations.

1070-986X/09/$25.00 © 2009 IEEE

Published by the IEEE Computer Society

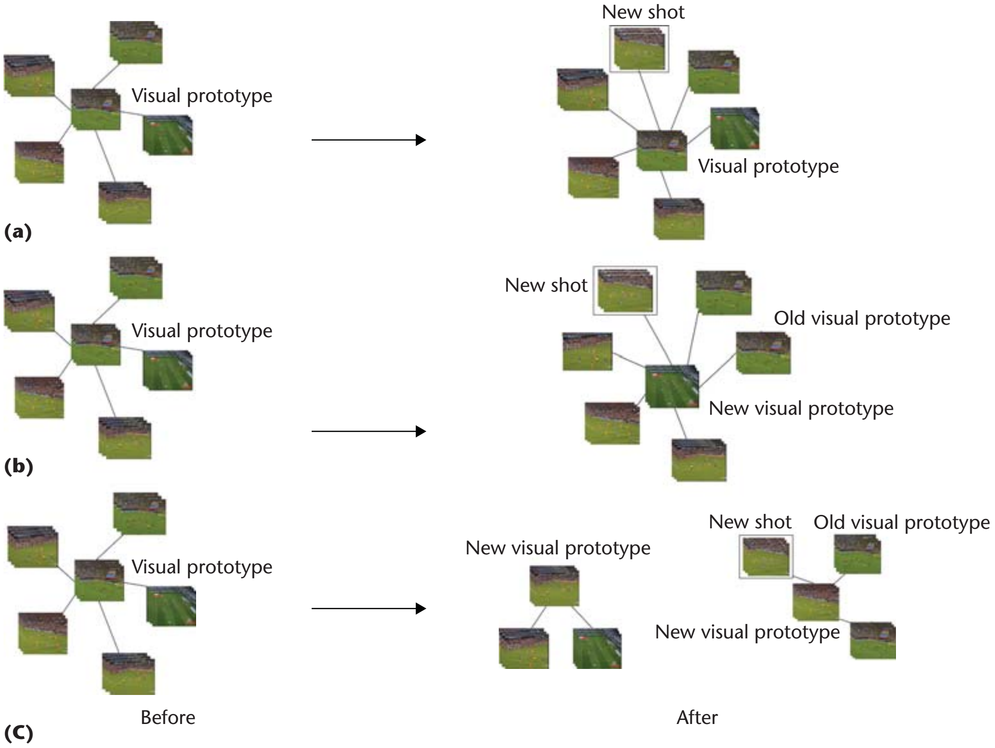

Figure 1. When the system assigns a new shot to a concept, it updates all the clusters of the concept. Some possible outcomes of the update include (a) the visual prototype of the cluster remains the same, (b) a new prototype is selected, and (c) the cluster is split and two new visual prototypes are selected.

Automatic video annotation framework

Our framework employs the Dynamic Pictorially Enriched Ontology model to perform video annotation and uses cluster updating to support temporal evolution of the visual prototypes. In this model, the ontology contains the linguistic domain concepts, their relationships, and visual instances. Concepts selected for visual instances are those that change in shape, appearance, and motion in their spatiotemporal pattern. In the ontologies used in the experiments, these concepts consist of sport highlights such as ''shot on goal'' and ''pit stop,'' or views such as ''long range'' and ''close up.''

Visual instances associated with the schema concepts include object identifier, visual descriptors, time label, and link to the raw data. We create these instances as the result of the matching between the descriptors of the raw visual data and the descriptors of one reference instance in the ontology. We use clustering to group instances that have some similarity in their visual or spatiotemporal patterns. In general, there are several clusters for each concept. Each cluster roughly corresponds to one concept modality in which that concept manifests itself in reality. To reduce the cost of descriptor matching, we define a visual prototype for each cluster to represent all the instances in the cluster. The median element in each cluster becomes the visual prototype, which is initially created using a training set of annotated data.

A special cluster, called the *unknown concept cluster*, includes the instances that have not been assigned yet to a cluster. Because visual instances might have a large number of modes in which they appear, our system considers any new instance to be new knowledge for the ontology. Every time a new instance is associated with a concept, the system updates all the clusters of that concept along with the unknown concept cluster. In this way, we can assign previously labeled unknown instances to some cluster or create new clusters as a result of the new instance. Ultimately, this process permits the system to effectively represent the instances' various appearances and motion patterns and provides a form of temporal evolution for the knowledge in the ontology. Figure 1 shows an example of clustering updates.

Visual instances are obtained by video segmentation and feature extraction. A shot-detection algorithm, called Linear Transition Detection (LTD), performs the video segmentation.

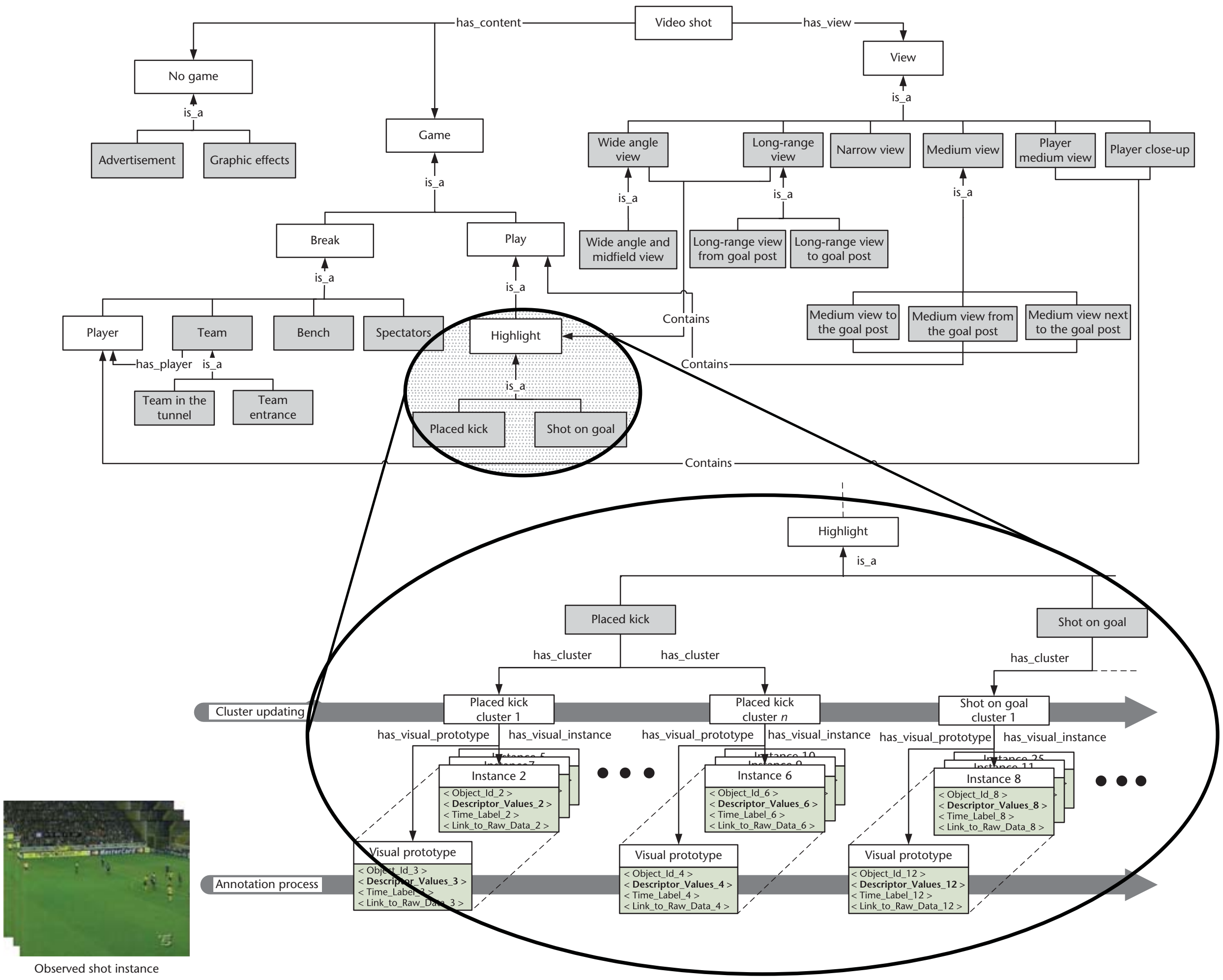

Figure 2. The video annotation framework with Dynamic Pictorially Enriched Ontologies showing a partial schema of the soccer video ontology with linguistic concepts and visual instances; instance clusters with visual prototypes; annotation that matches the observed instance with visual prototypes; and cluster updating.

By approximating a linear transition model, LTD is quite robust when it comes to detecting cuts, dissolves, and fades.[5] The descriptors extracted from video shots are of two types: low-level features and medium-level features. Examples of low-level features include color histograms or edge maps related to concepts used in different domains, such as scene setting or shot type. Examples of medium-level features include those that indicate a particular application domain, such as a playfield position in sport videos, or a specific entity, such as a face obtained from face-detection software. We use literals to represent the sequences of the descriptor values (computed from each frame of the shot) to account for different shot lengths and temporal order.

To test the system's capability, we used Formula 1 and soccer videos. Figure 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, while the *has_visual_prototype* property identifies the cluster prototype. In addition, the annotation and cluster-updating processes are schematized. The annotation process entails matching observed data instances with visual prototypes and consequent high-level concept association, while the cluster-updating process entails determining reclustering of the ontology's clusters from the observed data instances.

The system uses different clustering methods. For concepts where temporal ordering of descriptors is not relevant—typically because they refer to scene views such as close-up or wide-angle—we used Complete Link hierarchical clustering. We used the Mallow's distance $d(S_x; S_y)$ as a measure of the dissimilarity between the features vector of two shots S to account for shots of different length. The system assigns every shot to a different cluster, then proceeds with an iterative grouping of the two most similar clusters until it obtains a single cluster that contains all the instances. Each level of the hierarchy can be identified by its number n of clusters. The process obtains an optimal number $\tilde{n}$ of clusters automatically by maximizing a clustering score CS_n. To this end, we define the diameter $\Delta(W_i)$ of a cluster W_i and the distance $\delta(W_i; W_j)$ between two clusters W_i and W_j as:

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

And we define the clustering score at level n 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.

For concepts where the temporal evolution of descriptors is extremely relevant—for example, in a soccer "shot on goal" action, where player motion and play field zone values are important elements in distinguishing one action from the other—we used the 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 also accounts for the temporal order of the features. We obtain the distance, the sum of all the normalized Needleman-Wunch distances between the distinct components of the content descriptors, 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.[6]

Certain concepts occurring in a video can't be detected only from observing visual features. In some cases, these concepts can be recognized through analyzing the context and the content of the preceding and following shots. For example, in the soccer domain, some placed kicks that are not recognized using visual features can be recognized in that they are frequently preceded by player close-ups or medium-view shots. We can therefore define patterns that use temporal relations between concepts to improve shot annotation.

In our framework, we use rules to model these patterns and rule-based reasoning to recognize them. 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. We use SWRL to define the rules that model these patterns. With respect to Web Ontology Language axioms, SWRL permits us to express rules explicitly through if-then expressions, and provides built-in mathematical, logical, string comparison, and temporal operators, making it easier, even for nonprogrammer domain experts, to define and modify rules and rule constraints.[7]

Experimental results

We tested our annotation framework on the Formula 1 and soccer domains to show its general applicability and achievable performance improvements. Table 1 (next page) lists the concepts with visual prototypes from the two domains, representing dynamic actions, highlights, and common views used to show an overview of an event or its context. For the description of the visual content, we used low-level features, namely color layout, scalable color, edge histogram, and motion activity. For the soccer domain, we added a few medium-level features, namely the main camera motion direction and intensity, the framed playfield zone, and the number of visible players in the upper and lower part of the playfield. We described the "shot on goal" and "placed kick" highlights explicitly, considering the temporal evolution of their visual features.

We conducted the first experiment to check the performance of the video annotations, dependant on the number of shot instances in a training set used to create the visual prototypes, 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 used six distinct collections of video sequences extracted from the 2006 soccer world

Table 1. List of concepts with visual prototypes used in the Formula 1 and soccer ontologies. Indentations indicate that a concept is a specialization of another concept.

Domain	Concept	Description
Soccer	Wide angle view	Wide-angle view framed by the main camera
Soccer	Wide angle midfield view	Wide-angle view of the midfield area
Soccer	Long-range view	Long-range view focusing over middle area
Soccer	Long-range view from goal post	Long-range view as taken from the goal posts
Soccer	Long-range view to the goal post	Long-range view taking the goal post in the center
Soccer	Narrow view	Narrow-angle view as taken from handheld video camera
Soccer	Medium view	Players are fully displayed in the playfield
Soccer	Medium view from the goal post	Medium view as taken from the goal post
Soccer	Medium view to the goal post	Medium view of the goal post
Soccer	Medium view next to the goal post	Medium view where the goal post is lateral in the image
Soccer	Players medium view	Players view framed from a side camera near the playfield
Soccer	Bench	Coach and team staff view
Soccer	Player close-up	Players close-up view
Soccer	Team	View of the team
Soccer	Team entrance	View of the team entrance in the playfield
Soccer	Team in the tunnel	View of players taken in the tunnel before and after the game
Soccer	Spectators	View displaying supporters and cheering crowd
Soccer	Advertisement	View showing advertisement
Soccer	Graphic effects	View displaying any other elements such as computer graphics
Soccer	Shot on goal	Action where a player kicks the ball to the opponent's goal post to score a goal
Soccer	Placed kick	Action including penalty, corner, and free kick near the goal post
Formula 1	Wide angle view	Wide-angle view of the race track
Formula 1	Medium view	Medium view of the track
Formula 1	Car close-up	Car close-up view
Formula 1	Box staff	Staff view
Formula 1	Spectators	View displaying supporters and cheering crowd
Formula 1	Advertisement	View showing advertisement
Formula 1	Car-camera driver view	View of the driver from the camera car
Formula 1	Car-camera front view	View of the front of the car from the camera car
Formula 1	Box pit stop	View displaying a car and the team during the pit stop
Formula 1	Box car entry	Action where a car is entering the box
Formula 1	Box car exit	Action where a car is exiting from the box
Formula 1	Race start	Action at the start of the race

championship. The video collections contain different games, athletes, stadiums, and edit effects, and are different lengths. Each collection includes several concepts, such as wide-angle view, long-range view, medium view, spectators, team, and player close-up.

Table 2 lists the results of the experiment for this set in terms of average precision and recall. Initially, as listed in the first row in Table 2, we presented 81 shots to the ontology (trained with 25 shots annotated manually to create the initial set of visual prototypes). The system annotated 81 shots exploiting the initial set of visual prototypes. We manually annotated the concepts that were not already represented (11 concepts altogether). We repeated this process in several steps, with the addition of new shots at every step. As we presented new shots to the system, it updated the visual prototypes in the ontology. At step five, we used 1,158 shots containing 19 concepts. The quality of annotation improved from rows two to five.

The second set of concepts includes ''shot on goal'' and ''placed kick'' (rows six through eight) There is a great deal of variety in the visual appearance of these concepts, making a large number of prototypes necessary for effective annotation. Initially, we presented 68 shots

Table 2. Evaluation of annotation performance.

Video collection	Temporal feature ordering	Number of shots	Number of concepts	Shots used for ontology training	Average precision	Average recall
1	No	81	11	25	0.19	0.20
2	No	206	17	106	0.34	0.28
3	No	255	19	312	0.47	0.38
4	No	591	19	567	0.43	0.46
5	No	341	19	1158	0.55	0.52
6	Yes	68	2	50	0.43	0.27
7	Yes	68	2	100	0.53	0.40
8	Yes	68	2	150	0.62	0.60

to the system ontology that we had manually trained with 30 ''shot on goal'' and 20 ''placed kick'' shots (row six). In each of the following steps (row seven and eight), we used 50 additional shots for training. The experiment indicates that increasing the training set for the creation of the visual prototypes results in improved performance. In particular, the improvement in 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 the Dynamic Pictorially Enriched Ontologies model to capture the temporal evolution of the visual prototypes. To this end, we used the already annotated ''placed kick'' shots used for training in the previous experiment. Because they contain events filmed in the years 2001, 2005, and 2006, we inserted them into the ontology in three distinct steps so the clusters could be updated at each stage and the visual prototypes redefined. While feeding the ontology with this data, we kept track of the way in which the visual prototypes changed.

In particular, at each step, we registered the mean and variance of the shifts of the cluster centers with respect to their position in the previous step, together with the mean radius of each cluster and the number of clusters. In the case of cluster splitting, we calculated the shift of the original cluster center with respect to the closest of the new cluster centers. Table 3 shows the evolution of these parameters for each step. We calculated distances using the Needleman-Wunch distance defined previously. Doing so provided evidence of the different modes in which ''placed kick'' shots have been shown on TV from 2001 to 2006.

The ''placed kick'' shots from 2005 shifted the 2001 ''placed kick'' cluster centers, with an increase of both the mean cluster radius and the number of clusters. Adding the 2006 ''placed kick'' shots resulted in an increase in the number of visual prototypes, but a smaller average shift of the cluster centers and a slight decrease of the mean cluster radius. Indeed, these shifts reflect the fact that camera shots for soccer have changed considerably since 2005. The long phase of preparing for the kick (placing the ball, waiting for the placement of the opponents, and so forth) is now rarely shown, and has been replaced by player close-ups and medium views of the playfield, to display a faster and more dynamic scene.

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

Table 3. Dynamic evolution of ''placed kick'' visual prototypes (2001 to 2006).

Video collection	Years	Mean shift	σ^2 shift	Mean cluster radius	Number of visual prototypes
1	2001	n/a	n/a	3.6	4
2	2001, 2005	3.3	1.3	4.5	6
3	2001, 2005, 2006	1.0	1.7	4.4	8

Table 4. Annotation performance for the concepts defined in Table 1.

Domain	Concept	Correct	Unknown	Miss	False	Precision	Recall
Soccer	Wide-angle view	26	0	5	14	0.65	0.84
Soccer	Wide-angle midfield view	1	0	2	3	0.25	0.33
Soccer	Long-range view	12	4	9	9	0.57	0.48
Soccer	Long-range view from goal post	4	0	5	1	0.80	0.44
Soccer	Long-range view to the goal post	1	0	2	3	0.25	0.33
Soccer	Narrow view	2	1	7	5	0.29	0.20
Soccer	Medium view	26	2	11	18	0.59	0.66
Soccer	Medium view from the goal post	1	0	3	3	0.25	0.25
Soccer	Medium view to the goal post	6	1	2	5	0.54	0.67
Soccer	Medium view next to the goal post	1	2	4	1	0.50	0.14
Soccer	Players medium view	1	1	4	5	0.16	0.17
Soccer	Bench	7	1	14	3	0.70	0.32
Soccer	Player close-up	72	5	27	20	0.78	0.69
Soccer	Team	10	0	4	1	0.90	0.71
Soccer	Team entrance	1	0	3	1	0.50	0.25
Soccer	Team in the tunnel	1	1	1	4	0.20	0.33
Soccer	Spectators	15	0	3	12	0.55	0.83
Soccer	Advertisement	5	0	1	0	1	0.83
Soccer	Graphic effects	22	1	1	0	1	0.92
Soccer	Shot on goal	20	5	5	6	0.77	0.67
Soccer	Placed kick	11	9	1	12	0.48	0.52
Formula 1	Wide angle view	53	2	33	10	0.84	0.60
Formula 1	Medium view	47	4	31	46	0.50	0.57
Formula 1	Car close-up	76	4	18	54	0.58	0.77
Formula 1	Box staff	23	0	36	13	0.64	0.39
Formula 1	Spectators	41	0	43	13	0.76	0.49
Formula 1	Advertisement	96	3	4	8	0.92	0.93
Formula 1	Car-camera driver view	97	3	2	3	0.97	0.95
Formula 1	Car-camera front view	7	2	0	1	0.87	0.77
Formula 1	Box pit stop	57	0	44	29	0.66	0.56
Formula 1	Box car entry	78	0	21	20	0.80	0.79
Formula 1	Box car exit	81	3	19	54	0.60	0.79

Formula 1 and soccer domains, Table 4 lists the results. As the table illustrates, some concepts are poorly represented mainly due to the fact that they vary in appearance so much. We found that most false detections were caused by shots that have a narrow-angle view and medium-view concepts. In these cases, the system would need additional semantic information to recognize the concept properly.

The ''shot on goal'' concept performed well in terms of precision and recall. The low recall and large number of unknowns for the ''placed kick'' concept were caused by the dataset including shots filmed with different modalities. The car-camera concepts had excellent precision and recall because the shots could be distinguished easily from the other concepts. In addition, the ''box car entry'' and ''box car exit'' performed well because of the motion feature, while the ''box staff'' and ''box pit stop'' shots were easily confused with each other because of no unique motion characterization.

In the fourth experiment, we provided some evidence of the precision improvement and recall that is achievable with rule-based ontology reasoning, even with the addition of few simple rules. We performed the analysis for a few highlights of soccer and Formula 1 using rule-based reasoning with the Jess reasoning engine. We defined SWRL rule patterns for ''placed kick,'' "shot on goal," "box car exit," and "race start" as follows:

- Placed kick: 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" shot.
- 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" shot.
- 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" shot.
- 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" shot.

Figure 3 shows an example with the SWRL code for the "placed kick" pattern.

Table 5 shows the improvement obtained by rule-based reasoning. For soccer, the SWRL rules improve the recall, in particular. But for "box car exit," recall and precision remain almost the same because the number of shots classified as unknown are already low. The results indicate that SWRL reasoning can be useful in detecting concepts that are characterized by some temporal structure, such as the race start. Table 6 compares the performance obtained through the use of SWRL rules and ontology reasoning to the traditionally employed support vector machine (SVM) classification, for "shot on goal" and "placed kick" highlights.

To have a fair comparison, we trained the SVM classifiers (with a radial-basis-function kernel) on the same training set we used for the ontology, and represented video clips with the same vectors used for concept clustering in the ontology and with a fixed number of samples (five) per clip, so as to have feature vectors of the same length. The improvement caused by the Dynamic Pictorially Enriched Ontology model is essentially due to the fact that, in contrast to SVM, SWRL rules permit including some contextual information, namely temporal constraints, that allows the system to disambiguate situations without having to rely exclusively on visual features.

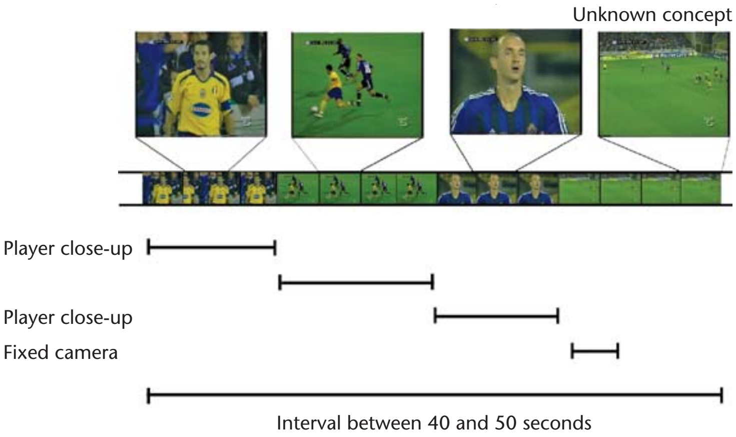

Figure 3. Example of a "placed kick" pattern.

Table 5. Precision and recall of concepts with visual prototypes (VP); with visual prototypes and SWRL reasoning (VP+SWRL), and with SWRL reasoning only.

	Shot on goal		Placed kick		Box car exit		Race start
Highlight	**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	0.67	0.77	0.52	0.86	0.79	0.81	0.70

Table 6. Precision and recall of concepts with visual prototypes and SWRL reasoning (VP+SWRL) and support vector machine (SVM) classifiers.

	Shot on goal		Placed kick	
Highlight	**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	0.77	0.60	0.86	0.71
Recall	0.77	0.60	0.86	0.71

Previous Work

Previous work has focused on the use of linguistic ontologies and appropriate classifiers to associate concepts with visual data. In these approaches, the ontology provides the conceptual view of the domain at the schema level while the classifiers observe the real-world sources, annotating the observed entities according to the nearest concept in the ontology. One work outlined a method to perform video annotation with the MediaMill 101 concept lexicon.[1] In this work, a computer trained classifiers to detect high-level concepts from low-level features, while using WordNet to derive high-level concepts relations to enhance the annotation performance. Another work defined an ontology to provide some structure to the Large Scale Concept Ontology for Multimedia lexicon, using pairwise correlations between concepts and hierarchical relationships to refine concept detection of support vector machine classifiers.[2]

Other works have postulated the idea of including visual data instances in the ontology to account for the variety of manifestations of visual information. These approaches apply feature detectors to raw data and match the extracted features to those of the concept instances in the ontology. One research team, in particular, defined a visual descriptors ontology, a multimedia structure ontology, and a domain ontology to perform video content annotation at the semantic level.[3] Another team included visual objects in the ontology instances, using qualitative attributes, such as color homogeneity, component distribution, and spatial relations, as descriptors.[4] Still another team proposed a visual concept ontology that includes texture, color, and spatial concepts and relations for object categorization.[5] Finally, in another work, researchers developed a framework for enhancing annotations by exploiting visual context and spatial relations.[6]

In the attempt to having richer annotations, other researchers have explored the use of reasoning over multimedia ontologies. In these cases, the works typically analyze spatiotemporal relationships between concept occurrences to distinguish between scenes and events and provide more fitting and comprehensive descriptions.[7,8] These works use inference to check relations and constraints that lead to consistent interpretation of image content, and sometimes use reasoning over whole sets of objects.

The inclusion of data instances in the ontology requires some mechanism for the management of the ontology evolution. In one project, researchers addressed the problem of temporal evolution of visual data by checking each visual instance to determine whether it could be associated with the existing abstract concepts or would require a new concept defined in the ontology.[9] In this project, researchers have proposed evolution patterns to define the kinds of action to be performed on the ontology.

In contrast, our own Dynamic Pictorially Enriched Ontology framework addresses the issues raised here in several ways: by including visual instances related to high-level concepts and identifying their spatiotemporal patterns; by defining visual prototypes and using them for automatic annotation; and by supporting the evolution of visual prototypes over time. Moreover, our approach not only emphasizes the need for spatiotemporal constraints among objects and entities for complex video-content interpretation, but also proposes the use of the Semantic Web Rule Language as an effective means to define, share, and refine rules that can lead to more effective concept definition and recognition.

References

1. C. Snoek et al., ''Adding Semantics to Detectors for Video Retrieval,'' *IEEE Trans. Multimedia,* vol. 9, no. 5, 2007, pp. 975-986.
2. Z.-J. Zha et al., ''Building a Comprehensive Ontology to Refine Video Concept Detection,'' *Proc. Multimedia Information Retrieval* (MIR), ACM Press, 2007, pp. 227-236.
3. S. Bloehdorn et al., ''Knowledge Representation for Semantic Multimedia Content Analysis and Reasoning,'' *Proc. European Workshop Integration of Knowledge, Semantics and Digital Media Technology* (EWIMT), 2004.
4. S. Dasiopoulou et al., ''Knowledge-Assisted Semantic Video Object Detection,'' *IEEE Trans. Circuits and Systems for Video Technology,* vol. 15, no. 10, 2005, pp. 1210-1224.
5. N. Maillot and M. Thonnat, ''Ontology Based Complex Object Recognition,'' *Image Vision Computing,* vol. 26, no. 1, 2008, pp. 102-113.
6. S. Dasiopoulou et al., *Semantic Multimedia and Ontologies Theory and Applications,* Springer, 2008.
7. S. Espinosa et al., ''Towards a Media Interpretation Framework for the Semantic Web,'' *Proc. Int'l Conf. Web Intelligence* (ICWI), IEEE Press, 2007, pp. 374-380.
8. B. Neumann and R. Moeller, ''On Scene Interpretation with Description Logics,'' *Cognitive Vision Systems: Sampling the Spectrum of Approaches,* LNCS, Springer, 2006, pp. 247-278.
9. S. Castano et al., ''Ontology Dynamics with Multimedia Information: The BOEMIE Evolution Methodology,'' *Proc. Int'l Workshop Ontology Dynamics* (IWOD), 2007, pp. 41-54.

Conclusions

The experiments we conducted in two different domains illustrate our model's effectiveness for video annotation, for supporting the temporal evolution of visual concepts, and for improving annotation performance through rule-based reasoning. Our future work will investigate techniques to incorporate automatic learning of set of rules to perform automatic video annotation. **MM**

Acknowledgments

This work was partially supported by the Information Society Technologies (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.
2. L. Hollink, G. Schreiber, and B. Wielinga, ''Patterns of Semantic Relations to Improve Image Content Search,'' *J. Web Semantics,* vol. 5, no. 3, 2007, pp. 195-203.
3. M. Naphade et al., ''Large-Scale Concept Ontology for Multimedia,'' *IEEE MultiMedia,* vol. 13, no. 3, 2006, pp. 86-91.
4. H. Gardner, *The Mind's New Science: A History of the Cognitive Revolution,* Basic Books, 1985.
5. C. Grana and R. Cucchiara, ''Linear transition detection as a unified shot detection approach,'' *IEEE Trans. Circuits and Systems for Video Technology,* vol. 17, no. 4, 2007, pp. 483-489.
6. M. Bertini et al., ''Video Annotation with Pictorially Enriched Ontologies,'' *Proc. Int'l Conf. Multimedia and Expo,* IEEE Press, 2005, pp. 1428-1431.
7. L. Hollink, S. Little, and J. Hunter, ''Evaluating the Application of Semantic Inferencing Rules to Image Annotation,'' *Proc. K-CAP Int'l Conf. Knowledge Capture,* 2005, pp. 91-98.

Marco Bertini is an assistant professor in the Department of Systems and Informatics at the University of Florence, Italy. His research interests include content-based indexing and retrieval of videos and Semantic Web technologies. Bertini has a PhD in electronic engineering from the University of Florence. Contact him at bertini@dsi.unifi.it.

Alberto Del Bimbo is a full professor of computer engineering at the University of Florence, Italy, where he is also the director of the Master in Multimedia Content Design. His research interests include pattern recognition, multimedia databases, and human–computer interaction. He has a laurea degree in Electronic Engineering from the University of Florence. Contact him at delbimbo@dsi.unifi.it.

Giuseppe Serra is a PhD student at the Visual Information and Media Lab at the Media Integration and Communication Center, University of Florence, Italy. His research interests include multiple-view geometry, self-calibration and 3D reconstruction, and video understanding based on statistical pattern recognition and ontologies. Serra has a laurea degree in computer engineering from the University of Florence. Contact him at serra@dsi.unifi.it.

Carlo Torniai is a postdoctoral research associate in the Department of Electrical Engineering at the University of Southern California. His research interests include the Semantic Web, knowledge management, and learning technologies. Torniai has a PhD in computer science from the University of Florence, Italy. Contact him at torniai@usc.edu.

Rita Cucchiara is a full professor at the University of Modena and Reggio Emilia where she heads the ImageLab laboratory (see http://imagelab.ing.unimore.it). Her research interests include pattern recognition and computer vision for video surveillance and multimedia. Cucchiara has a PhD in computer engineering from the University of Bologna, Italy. She is a Fellow of the International Association for Pattern Recognition (IAPR). Contact her at rita.cucchiara@unimore.it.

Costantino Grana is an assistant Professor at the University of Modena, Italy. His research interests include multimedia information analysis, focusing on image and video concept detection, and historical document analysis. Grana has a PhD in information engineering from the University of Modena and Reggio Emilia. Contact him at costantino.grana@unimore.it.

Roberto Vezzani is an assistant Professor in the Faculty of Mathematical, Physical, and Natural Sciences at the University of Modena and Reggio Emilia. His research interests include computer vision and pattern recognition for video surveillance systems. Vezzani has a PhD in information engineering from the University of Modena and Reggio Emilia. He is a member of the IEEE and GIRPR (IAPR, Italy). Contact him at roberto.vezzani@unimore.it.