Dynamic Pictorial Ontologies for Video Digital libraries Annotation

M.Bertini, A.Del Bimbo, C. Torniai
DSI
Università di Firenze, Italy
{bertini,delbimbo,torniai}@dsi.unifi.it

C.Grana, R.Cucchiara
DII
Università di Modena e Reggio Emilia, Italy
{rita.cucchiara,costantino.grana}@unimore.it

ABSTRACT
In this paper, we present the dynamic pictorial ontology paradigm for video annotation. Ontologies are often used to describe a given domain for different goals, including description of multimedia data. In the case of video annotation, the visual knowledge cannot be described using only abstract concepts but is more effectively represented in a visual form. To this aim, we introduce *visual concepts*, elicited from the data set as the most representative prototypes that specialize abstract concepts. The ontology created is intrinsically dynamic since it must embrace the perceptual and visual experience during annotation. Thus visual concepts can change, adapting to the multimedia content analyzed. Motivation for this new ontology paradigm are discussed together with a proposal of a framework for ontology creation, maintenance, and automatic annotation of video. The creation and usage of dynamic pictorial ontologies have been tested for soccer domain exploiting low level perceptual features and higher level domain features.

Categories and Subject Descriptors
H.3.7 **[Information Systems Applications]:** Digital Libraries

H.2.8 **[Database Applications]:** Image databases

General Terms
Video annotation.

Keywords
Video annotation, video retrieval, multimedia ontologies, multimedia extraction and annotation

1. INTRODUCTION
Managing knowledge in video digital libraries is a foundational problem for the new generations of content-based retrieval systems. Ontologies are commonly used for knowledge representation of different domains. For the video domain research activities have been focused on ontology definition [1,2,6], ontology standards and languages [3] and on methodologies to connect knowledge extracted from data to the concepts of the ontology [4]. Traditional ontologies are based on linguistic concepts and typically exploited in annotation and information retrieval of semi-structured XML-based documents. However, traditional ontologies are substantially inadequate to support efficient annotation and retrieval of video documents. In fact, concepts and categories expressed in linguistic terms are not rich enough to fully describe the diversity of the visual events that occur in a video sequence and cannot support video annotation and retrieval up to the level of detail of a pattern specification.

A simple example may help to understand this point. Let's consider attack action highlights in a video of a soccer game. The highlights that can be classified as attack actions have many different patterns. The patterns may differ each other by the playfield zone where the action takes place, the number of players involved, the players' motion direction, the speed and acceleration of the key player, etc. Although as spectators we are able to distinguish between attack actions, and cluster them into distinct classes, nevertheless, if we want to express each pattern in linguistic terms we should use a complex sentence, explaining the way in which the action was developed. The sentence indeed should express the translation of the user's visual memory of the action into a conceptual representation where concepts are concatenated according to spatio-temporal constraints. In this translation, some visual data will be lost (we typically make a synthesis that retains only the presumed most significant elements), some facts will not be reported and, probably most important, appropriate words to distinguish one pattern from the other may not be found.

Some early attempts to solve the inadequacy of traditional linguistic ontologies to support modeling of domain knowledge up to the level of pattern specification, have envisioned the need that video domain ontologies incorporate both conceptual and perceptual elements. The basic idea behind all these researches is that, although linguistic terms are appropriate to distinguish between broad event and object categories in generic domains, they are substantially inadequate when they must describe specific patterns of events or entities, like those that are represented in a video, and more in general in any perceptual media.

In [18], Jaimes et al. suggested to categorize the concepts that relate to perceptual facts into classes, using modal keywords, i.e. keywords that represent perceptual concepts in several categories, such as visual, aural, etc. Classification of the keywords was obtained automatically from speech recognition, queries or related text. They introduced the term "multimedia ontology" to represent ontologies that in addition to linguistic concepts also include visual and auditory concepts. In [19], perceptual knowledge is

MFMS'07, September 28, 2007, Augsburg, Bavaria, Germany.

instead discovered grouping previously annotated images into clusters, based on their visual and text features, and extracting semantic knowledge by disambiguating the senses of the words in the annotations with WordNet and image clusters. Visual prototypes instances are then manually linked to the domain ontology.

In [20,21], a linguistic ontology for news videos is used to find recommendations for the selection of sets of effective visual concepts. A taxonomy composed by almost 1000 concepts was defined providing a multimedia controlled vocabulary. The taxonomy was mapped into the Cyc knowledge base [22] and exported to OWL. Other authors have explored the possibility of extending ontologies so that also structural video information and visual data descriptors could be accounted for. In [5] Athanasiadis et al. implemented MPEG-7 visual descriptors (i.e. low-level general purpose features) linked to high level domain specific concepts. In this way traditional methodologies of ontology concept management were extended to audio and video concepts. In [23], a Visual Descriptors Ontology and a Multimedia Structure Ontology, respectively based on MPEG-7 Visual Descriptors and MPEG-7 Multimedia Description Schema, are used together with a domain ontology in order to support video content annotation. A similar approach was followed by [24] to describe sport events. In [25], a hierarchy of ontologies was defined for the representation of the results of video segmentation. Concepts were expressed using keywords of an object ontology: MPEG-7 low-level descriptors were mapped to intermediate level descriptors that identify spatio-temporal objects. In all these solutions, structural and media information are still represented through linguistic terms and fed manually to the ontology. In [26] three separate ontologies modeling the application domain, the visual data and the abstract concepts were used for the interpretation of video scenes. Automatically segmented image regions were modeled through low-level visual descriptors and associated to semantic concepts using manually labeled regions as a training set. Text information available in videos, obtained through automatic speech recognition and manual annotation, and visual features were automatically extracted and manually assigned to concepts or properties in the ontology in [4]. In [2], qualitative attributes that refer to perceptual properties like color homogeneity, low-level perceptual features like model components distribution, and spatial relations were included in the ontology. Semantic concepts of video objects were derived from color clustering and reasoning. In [29] the authors have presented video annotation and retrieval based on high-level concepts derived from machine learned concept detectors that exploit low level visual features. The ontology includes both semantic descriptions and structure of concepts and their lexical relationships, obtained from WordNet.

Other researchers have proposed integrated multimedia ontologies that incorporate both linguistic terms and visual or auditory data and their descriptors.

In [28] images are automatically segmented, and the areas obtained are compared to existing prototypical descriptions of concepts. The prototypes are defined statically, and are manually segmented and linked to MPEG-7 low-level descriptors. In [30] a taxonomy was defined for video retrieval, where visual concepts were modeled according to MPEG-7 descriptors. Video clips were manually classified according to taxonomy and unsupervised clustering was employed to cluster clips with similar visual content. In [13] we have proposed a multimedia ontology, referred to as "pictorially enriched ontology", where concepts with a visual counterpart (like entities, highlights or events) were modeled with both linguistic terms and perceptual media, like video and images. Perceptual media were associated with mid and high-level descriptors of their structure and appearance. A similar approach has been recently adopted within Boemie IST research project [27].

However, none of these works has taken into account the different behaviors of the linguistic and multimedia parts of the ontologies, in particular the fact that the visual concepts change in time, according to the experience and the new knowledge that is added.

In this paper we present a dynamic pictorial ontology that can be used for video annotation and retrieval by content, where concepts with a visual counterpart (like entities, highlights or events) have been modeled with both linguistic terms and perceptual media, as video and images. Perceptual media are associated with descriptors of their structure and appearance. Visual concepts are defined by means of selected prototypes that represent a set of similar instances. They dynamically change as new knowledge is added to the ontology.

Features that model visual concepts can be either low-level, i.e. basic perceptual features used for recognition of basic content features, or high-level, i.e. semantic features that can be used for domain-specific content recognition. Both approaches can be explored separately, depending on the effectiveness/generality tradeoff, or intermingled in a single framework. Examples are reported for soccer video. The main goal of the tests is to prove how the dynamic concepts can evolve while the annotation is growing enriching the base of knowledge and thus improving the possibility to provide a correct clip annotation. The paper is organized as follows: in the next section some motivations for dynamic enriched ontologies are discussed. Section 3 presents the structure of both the proposed framework and the ontology. Then section 4 discusses usage of pictorial ontology for soccer domain using both low-level and high-level semantic features together with experimental results. Conclusions and future work are presented in Section 5.

2. MOTIVATION FOR DYNAMIC PICTORIAL ONTOLOGIES

Philosophy has addressed the definition and study of ontologies, much before than computer science. In particular, in philosophy a clear distinction is made between ontology and epistemology [8]. The term "Ontology" is concerned with "what is said to exist in some world, which potentially can be talked about" or sometimes with "a set of terms and their associated definitions intended to describe the world in question". The term "Epistemology" denotes "the nature of human knowledge and understanding that can possibly be acquired through different types of inquiry and alternative methods of investigation", or "what is the nature of the relationship between the knower and what can be known".

Computer scientists usually merge the two concepts of ontology and epistemology in a single operating framework [7]. They intend ontology as a set of terms and their associated definitions that can be used to describe the world, according to the view of the knower.

As ontologies have evolved to include descriptors of images and other media, additional questions have been raised related to how perceptual media link to linguistic concepts, what they actually represent and in which way they should be managed. First of all, linguistic and perceptual concepts have different nature. We can refer to the dispute between the Platonic and Aristotelic views of concepts. In the Platonic theory, ideas embed permanent, self-contained, objective items of exact knowledge. In the Aristotelic view, concepts are not abstractions but mere duplicates of things observed in reality, and correlated to matter; concepts are therefore dependent on subjective and temporal human experience, and hence subjected to changes with time and individual perceptions.

According to this, we should regard traditional linguistic ontologies as static descriptions of application domains (considering them as implementations of the Platonic theory of ideas), while ontologies that include perceptual concepts should instead be regarded as the implementation of the Aristotelic view. In fact, while still keeping a clear separation between concepts and instances, these ontologies permit to link concepts to their real manifestations. Descriptors of perceptual features, linked as specializations of abstract concepts permit to establish a relationship between concepts and the perception of their instances in the reality. The identification of the differences in which the same abstract concept manifests in reality, can be performed by means of clustering of perceptual concepts according to the similarity of their descriptors. Through this fundamental process the centers of the obtained clusters can be defined as perceptual concepts prototypes.

Moreover, since perceptual concepts are real instances of abstract concepts, perceptual concept prototypes should adjust themselves to changes and modifications that may occur through time. In other words, the occurrence of new instances adds new knowledge and consequently modifies the prototypes, therefore establishing the dependency of perceptual concepts prototypes on experience.

It is important to notice that perceptual concept prototypes that are introduced in these ontologies are not meta-concepts or structural description of the reality but an immediate visualization of a concept. This has a direct relationship with Wittgenstein's *picture theory* [10], according to which "world consists entirely of facts" and human beings are aware of the facts by virtue of their mental representations which are most fruitfully understood as *picturing* the way things are ("we make to ourselves pictures of facts" and "the picture is a model of reality").

Dynamic pictorial ontologies as defined in our approach answer both to ontological and epistemological questions and implement all the requirements arisen from the inclusion of perceptual concepts. In fact visual concepts, while belonging to a different abstraction level with respect to linguistic concepts, at the same time constitute a way in which those concepts manifest, can be fully understood and immediately recognized.

There are three key properties of the proposed ontologies:

1. They embed a model of reality that is typical of video domain.
2. Perceptual concepts directly represent the way in which things, facts or events manifest.
3. Perceptual concepts are not defined a priori but learned depending on the experience and context. Therefore they dynamically change as the reality depicted in the video changes.

3. IMPLEMENTING VISUAL CONCEPTS

A picture of the framework for creating and maintaining dynamic pictorial ontologies and exploiting them for annotation is given in Figure 1.

Since the input multimedia footage is composed by raw or edited video, the first necessary step is the structural annotation. This step is composed by a task for partitioning data in elementary units (usually shots) and a task for metadata extraction. Visual features that will be used for the creation and maintenance of the ontology and for the automatic annotation of videos must be computed from each shot.

Features can be classified as:

i) Low-level perceptual features, i.e. generic visual descriptors such as color histograms, edge maps, etc. that are usually related to generic concepts such as scene settings, shot type, classification of indoor/outdoor scene, etc.

ii) High-level semantic features, i.e. high level descriptors such as face detection and recognition, superimposed text, etc. and specific descriptors that are usually related to the domain such as playfield detection and recognition in sports videos, anchorman detection in news and target tracking in video surveillance.

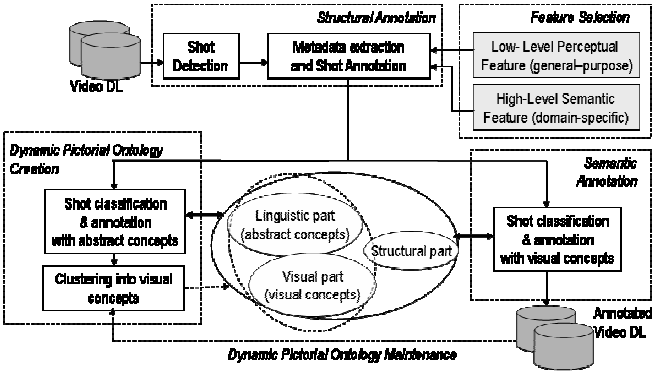

Figure 1 - Annotation and Retrieval Framework Overview

The Dynamic Pictorial Ontology is composed of two main parts, a structural part and a domain part. The latter is composed by:

1. A linguistic part containing abstract concepts expressed using linguistic terms
2. A visual part containing visual concepts

In the initial step of ontology creation the domain ontology is defined by domain experts and expressed in linguistic terms. The pictorial ontology is created automatically with an initial step of machine learning with training video by associating video samples to linguistic concepts.

Then visual concepts are defined automatically with a visual clustering process; this process elicits from the set of clips

associated to each linguistic concepts the subset that can be considered as visual concepts. This pictorial representation of the concepts requires to extract the subset with the minimum cardinality as possible and with the maximum visual diversity. It is automatically expanded dynamically, in the ontology maintenance and expansion step.

Automatic annotation using the dynamic pictorial ontology is performed with the evaluation of visual similarity (a typical paradigm of content-based retrieval). Perceptual or semantic features can be exploited depending on the tradeoff between efficacy/generality. New clips are associated to a visual concept of the pictorial ontology and inherit the linked linguistic concepts. At the end of an annotation phase a re-clustering process is exploited for the dynamic ontology maintenance. The concepts and their cardinality can change by changing the center and the number of the clusters. In such a manner, the knowledge is enriched by the experience.

In Figure 2 a schema for a generic dynamic pictorial ontologies, which is defined using OWL standard, is shown for a generic video domain. Both the Domain ontology and the Video structure ontology are depicted..

The latter is used during the structural annotation process described before. The main goal of this structure ontology is to provide the desired level of granularity for annotation and retrieval purposes. This is achieved by linking the structural instances (shots, clips, image regions, etc.) to the visual prototypes.

The Domain Ontology contains all the concepts and relations that define the domain being analyzed. It includes the high level concepts (expressed in linguistic terms) that name the entities and facts of the domain, and the perceptual elements that model the visual patterns of the specializations of the concepts of the linguistic part. The visual prototypes are duplicates of facts or events observed in reality. To account for the many perceptually different patterns in which they can manifest, visual prototypes are clustered according to the similarity of their spatio-temporal patterns. For each cluster, a visual prototype is obtained as the representative element for that pattern. In the ontology architecture proposed, the visual prototypes act as a bridge between the domain ontology and the video structure ontology. A visual prototype is a structural element of the video (either a clip, a shot, a frame or part of a frame) and at the same time is linked to a linguistic concept of the domain ontology. This link plays an important role for automatic annotation of new video sequences and for retrieval by content.

4. EXAMPLES FOR SOCCER VIDEO

This paradigm of dynamic pictorial ontology has been defined in the VAPEON (Video Annotation with Pictorial Enriched Ontology) project of the Network of Excellence in Digital Library DELOS (EC, IST Program VI FP). We applied it in different contexts of sport video DLs, exploring different approaches with different granularity and abstraction levels of the visual features.

Together with low level features, higher level features can be used to create the pictorial ontology (adding visual concepts for higher level and domain specific linguistic concepts) and to refine the annotation provided by using perceptual features only. In fact, while general purpose features can be useful to achieve a classification of the video content based on the more general concept of the ontology (such as setting, framing, type of the shot), higher level features can specialize this categorization according to the domain specific concepts.

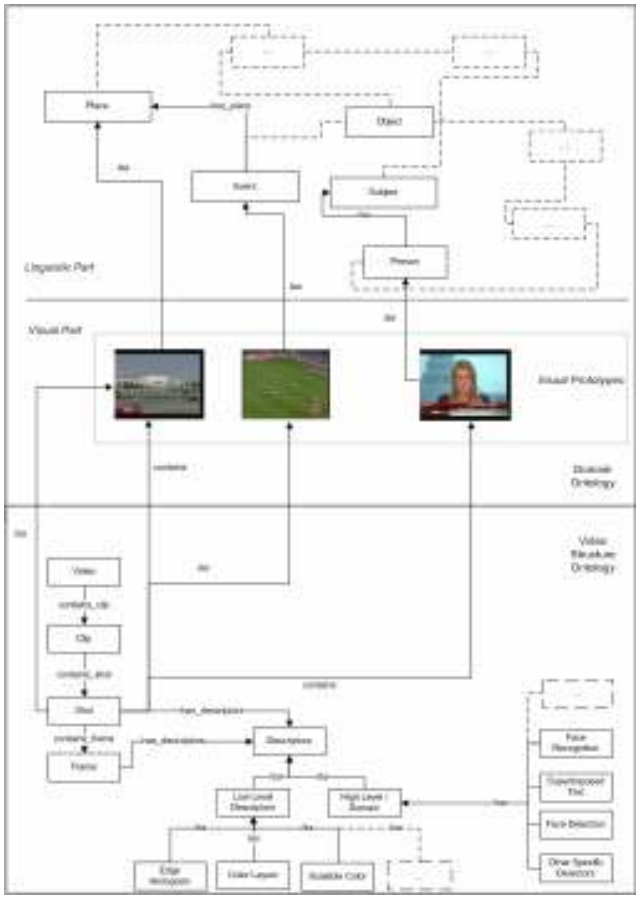

Figure 2 - Generic Dynamic Pictorial Ontology Schema

For example in soccer domain, with low level features it is possible to recognize the various types of shots framing actions but, in order to recognize the type of action, higher semantic level features have to be used.

The creation and usage of dynamic pictorial ontologies have been tested for soccer domain exploiting both the low level perceptual features and higher level domain features. In the following usage of clustering for visual prototypes creation and ontology extension together with annotation mechanism are presented along with experimental results.

The soccer domain ontology has been created and tested using many videos of the Italian World championship 2006, published by "The Gazzetta dello Sport". A subset of the automatically detected clips (about an hour of video) annotated with an ontology of 19 concepts is available at the web site http://imagelab.ing.unimo.it.

4.1 Visual Structure Processing

When analyzing broadcast or edited video, segmentation of video in shots or, if necessary, in sub-shots is a common pre-requisite. We developed a shot detection algorithm called LTD (Linear

Transition Detection) devised for typical linear transitions such as cuts or dissolves; it has proven to be sufficiently robust also in some not linear transitions typically adopted in broadcast TV [12]. Shot detection algorithms have been developed and tested in TRECVID 2005 [11,12].

Shots and transitions are described according to the MPEG-7 standard. Non uniform shots are divided into sub-shots, by clustering frames with perceptual similarity based on color and motion and the resulting clips are annotated in MPEG-7.

The shots and clips resulting from the segmentation process are used to build the dynamic pictorial ontology.

4.2 Perceptual features for Dynamic Pictorial Ontology

Most of the research activity in content-based retrieval addresses perceptual features, general-purpose enough to be applied in different contexts and sufficiently selective to distinguish different perceptual patterns. Of course, these features are not sufficiently specialized to recognize peculiar concepts of a domain. A possible compromise consists in the use of low level features in conjunction with an initial supervised training step. The atomic visual units are the clips, either shots or sub-shots, typically of a variable length with tens or hundreds of frames. The linguistic ontology is defined as a hierarchical taxonomy with abstract concepts of the domain. For soccer videos, an ontology of 30 perceptual concepts has been defined, describing the point of view, the general content and the type of video taken. In the highest level of the hierarchy *in-field* and *out-field* concepts distinguish the field of view. The former is specialized into *long, close, medium*, views. They are further specialized into other concepts until the final leaves that are 19 concepts such as "*long shots in front of the goal*", "*kickoff*", "*corners*" etc. The latter is specialized into concepts such as "*entry in the playing field*", "*spectators*", "*newspaper*", etc.

Each clip is described by a set of low level features. Initially standard MEPG-7 features for color, motion and texture have been explored to be fully compliant with the standard. In particular the Scalable Color Descriptor was selected, which is a color histogram in the HSV color space, with 16 values in H and 4 values in each S and V (256 bins in total). Then other features have been adopted, such as the Enhanced HSV histogram with achromatic point [14]. Features are extracted for a variable number of frames selected with a clustering process. The resulting feature vector has a variable length since can be composed by a set of features with variable number of frames. For this reason a non-trivial similarity measure must be adopted taking into account the different vector length. In this case, the *Mallow distance* has been exploited, as initially proposed in [15], specifically defined for clip similarity.

According to the dynamic pictorial ontology paradigm, an initial training step is performed with a manual annotation of shots against the aforementioned abstract concepts. Then the sets of clips associated with each concept are clustered into a variable set of clusters. We adopt a hierarchical clustering method, based on *Complete Link*. Starting from the separate set of elements (the clips) the Complete Link group them recursively by creating a tree; the process ends when the root is reached, i.e. a single group that contains all the elements. By fixing the number of levels, a number of clusters can be fixed. Conversely, an automatic selection of the number of cluster can be provided for instance with the Dunn's Separation Index. Better results have been obtained with the following approach that is based on a measure of distance between the W_i *and* W_j clusters. Being $d(S_x,S_y)$ the Mallow distance between the feature vector of two S shots, we define the cluster distance as

$$\delta(W_i,W_j) = \min_{S_x \in W_i, S_y \in W_j} d\left(S_x,S_y\right). \quad (1)$$

The *Clustering Score* at level *n* is defined as

$$CS_n = \min\left(\Delta_1 - \Delta_n, \delta_n\right) \quad (2)$$

where

$$\begin{aligned} \Delta_n &= \max_{W_i \in E_n} \Delta\left(W_i,W_i\right) \\ \delta_n &= \min_{W_i,W_j \in E_n, i \neq j} \delta\left(W_i,W_j\right) \end{aligned} \quad (3)$$

The selected level is the one which maximizes CS_n. It is possible to observe that Δ_n and δ_n are both monotonically increasing with *n*, thus CS_n has a unique global maximum. Therefore, the clustering algorithm can be stopped when CS_n start to decrease without computing the remaining levels. Then a single prototype, elicited as pictorial concept, can be generated from each cluster, by computing the *M* average feature vectors. The clip which minimizes the distance from the prototype features is associated to it, in order to provide a representative of the pictorial concept.

Then the automatic annotation of clips is performed associating them to the nearest neighbor pictorial concept, using the same low level features and similarity measure. Annotated clips can be further exploit to update the ontology in a dynamic maintenance step, by running the hierarchical clustering again and thus selecting new pictorial concepts, possibly more representative of the visual manifold of each abstract concept.

In Figure 3 the PEANO[1] interface if shown for the initial training phase. With a friendly drag-and-drop interface, clips initially belonging to an *Unknown* class can be associated with abstract concepts. In the example the clip that shows the Italian coach Lippi is associated to the concept close-ups and added to the set

Figure 3 - PEANO Interface

of prototypes of that concept. Then the set of hundreds of prototypes is clustered in a limited set of some units or tens of pictorial concepts, using the process described.

The clustering results are measured using the detection rate (the sum of the elements in the diagonal of the confusion matrix). They depend on the size of the initial training set, the type of features and the number of type of concepts in the ontology. While the knowledge base increases the annotation quality increases too. We tested how the dynamic maintenance improves the quality with the six experiments described in Table 1. Six collections of clips of the edited videos of the world-championship 2006 are used without any unfair clip selection. They contain very different clips of different games, in different stadium, with different players and special effects. In the table the length and the number of clips of each video is shown. In each test we used a set of clips to create the pictorial concepts as indicated in the fourth column. In the first example we manually annotated 12 clips only within the subset of 25. None of the other 13 clips was correctly annotated. Then we used all the set of 25 clips of the first video to create the pictorial concepts for the second one, achieving a detection rate of 30.8%. At each step, all the clips of the previous step (with a manual correction when errors occurred) are added. In the last test, 1158 clips, manually verified against ground truth, are automatically clustered into a variable number ranging between 10-25 for each of the 19 abstract concepts. About 60% of the new clips are correctly classified. This is a very interesting result that has been further explored, by changing the abstract concepts in the ontology and the type of features. For instance, by decreasing the number of concepts in the hierarchical taxonomy, the detection rate increases considerably: it becomes 67.74% with 11 concepts and 87.94% with 4 concepts only. This allows to select with a sufficient reliability a subset of clips with a given perceptual view of the field that can be further explored with semantic domain specific features.

Table 1 - Results of video annotation using perceptual features

video	length	No.clips	Clips for the pict-en ontology	Detection rate
1	00:40	25	12	0%
2	09:05	81	25	30.8%
3	11:27	206	106	36.4%
4	13:30	255	312	41.9%
5	29:24	591	567	52.01%
6	18:37	341	1158	63.96%

4.3 Semantic features for Dynamic Pictorial Ontology

High level descriptors can be either specific to a single domain only, or sufficiently generic so that they can be shared across multiple domains. Examples of domain specific descriptors are playfield descriptors in sports, interview sequence detectors in news, object tracking descriptors for surveillance. Other descriptors such as face detectors or superimposed text detectors may be used in all the domains where people and information are depicted.

[1] For details: http://imagelab.ing.unimo.it/imagelab/peano.asp

In the following we will show the use of dynamic pictorial ontologies applied to the soccer domain, presenting domain specific descriptors that are related to the sport domain in general and to the soccer domain in particular. We are interested in modeling visual concepts for soccer highlights that are present in play scenes. They are distinguished on the basis of the spatio-temporal combination of a reduced set of visual features: the *camera motion direction and intensity* (approximately modeling the key players' motion); the *playfield zone*; the *number of players* in the upper and lower part of the playfield. The features are obtained from the compressed and un-compressed video domain, as described in [16]. Each clip is described using a set of feature vectors whose length is proportional to the length of the clip after a frame subsampling of the clip.

Unsupervised pattern clustering is the mechanism used to define and update the perceptual part of the ontology. Patterns in which perceptual facts manifest are distinguished each other by clustering their descriptors, and the centers of the clusters are assumed as pattern prototypes and inserted as pictorial concept in the ontology. We employ fuzzy c-means (FCM) clustering [17] to take into account that image regions or clips in a cluster, have some degree of similarity also to image regions or clips of different clusters; for instance some actions share a common pattern during parts of their development.

New patterns that are analyzed for annotation are considered as new knowledge of the context. Therefore, as they are annotated using the ontology, they can modify the clusters and their centers, thus dynamically changing the perceptual part of the ontology.

The soccer ontology is based on many abstract concepts starting from an initial classification of *scene* and *highlight* concepts that are further specialized. For scene and highlight clustering, we employ a distance function modeled as the sum of all the normalized *Needleman-Wunch distances* between the U components of the feature vectors so as to account that two clips c_i, c_j may have different duration and their feature values may exhibit different temporal changes. The distance function is defined as follows:

$$d(c_i, c_j) = \frac{\sum_U NW(U_{c_i}, U_{cj})}{\min(length(c_i), length(c_j))}$$

This allows to take into account the dynamic aspects of videos and the video editing operations typically used (e.g. trimming, insertion, filtering). The distance is a generalization of the *Levenshtein edit distance* and has been used since the cost of character substitutions is an arbitrary distance function [16].

The processing time needed to evaluate the edit distance is low due to the fact that the alphabet needed to express the domain specific descriptors is very limited, compared the alphabet that would be required for the generic low level descriptors.

The visual concepts of domain specific concepts related to high level descriptors are updated using a two steps algorithm.

In the first step, an initial classification is performed evaluating the distance between visual prototypes and each incoming clip. A clip is classified as a highlight or scene type if its distance from a visual prototype linked to this abstract concept is lesser than a computed threshold τ_1. In this step a special class (*Unknown scene/highlight class*), that includes all the clips that have not been assigned to some highlight or scene class, is created.

The second step analyzes each clip classified as *Unknown scene/highlight*. A clip is classified as type of scene or highlight if enough clips of the same type have a distance from the clip that is lesser than a computed threshold τ_2. If a clip is classified as an highlight or scene type then FCM clustering is performed to re-evaluate the corresponding visual prototypes.

τ_1 is computed as one half of the minimum of the distances between all the visual prototypes in the ontology; τ_2 is the average of the radius of all the clusters of one specific highlight; τ_1 must account for the possibility that the perceptual part of ontology could not include enough representative facts. τ_2 is less conservative because it exploits all the new knowledge of perceptual facts that has been added to the ontology.

Adding new knowledge (i.e. new video clips) induces changes in the clusters due to the fact that their elements change and, consequently, their cluster centers are redefined (i.e. the visual prototypes change). Details on the annotation system for soccer and the performance achieved have been described in [16]. We have analyzed the effects induced by the introduction of new highlight examples on the clustering of shot on goals and placed kick highlights. To this end, we used a test set of 90 shot on goal clips and 60 placed kick clips, extracted from 36 national and international matches, played in 2001, 2005 and 2006. For each year 30 shots on goal and 20 placed kicks were used. In particular, we analyzed the evolution of shot on goal and placed kick, from 2001 to 2006. The pictorial ontology was initially built with the highlights of year 2001. Then highlights of 2005 and 2006 have been added. At each step the mean shift between the clusters centers of the highlights, the variance of the shift, the mean clusters radius and the number of clusters have been recorded. Table 2 and Table 3 show the evolution of these parameters for the two highlights; on each row of the tables have been reported the changes w.r.t. the previous ontology status, with the exception of row one and four, that contain the ontology used as initialization.

In Table 2, it can be observed that, for shot on goal, from year 2001 to year 2005 (2001+2005) the mean cluster radius is decreased while the number of clusters has increased; the mean shift of cluster centers is relatively high. This combination of figures is an index of considerable changes in the ontology due to the introduction of new visual prototypes. The addition of the 2006 highlights introduces few changes to the previous knowledge. In fact, as can be observed, the 2006 shot on goal patterns look similar to those observed in 2005 (see row 5): when introduced in the previous clustering they determine a very small shift of the cluster prototypes and the introduction of few new prototypes. It is interesting to notice that these figures put into evidence the evolution of soccer play in shot on goal actions from 2001 to 2006. Since there are no visible changes in camera shooting (camera takes are always registered from the main camera, following the ongoing action and zooming on the goal box area at the end of the shot), the changes observed must be ascribed to the changes in the way in which the action is performed.

Table 2 - Evolution of the ontology knowledge on the position of the clusters centers for Shot on goal highlight 2001-2006

Years	Mean shift	σ^2 shift	Mean cluster radius	n. clusters
2001			5.51	4
2001+2005	6.42	2.45	3.03	12
2001+2005+2006	2.70	1.39	3.50	13
2005			1.32	9
2005+2006	1.10	0.40	2.64	10

A similar but less pronounced evolution is observed for placed kick highlights. For this highlight, the 2005 and 2006 patterns determine both a shift of 2001 cluster prototypes and a change in the number of clusters (see rows 2 and 3); in particular it is the introduction of 2005 patterns that causes the larger cluster centers shift and the introduction of new prototypes. The addition of the 2006 knowledge reinforces the trend of the changes, but the difference with the knowledge of year 2005 (row 4) is relatively small in terms of shift, as can be noted in row 5. Differently from the case of shot on goals, this clustering evolution is not due to a corresponding evolution of the way in which the action is played (placed kicks do not have substantial variations in the way in which they are played) but instead reflects a change in the way in which they are shooted. Visual inspection of the videos has confirmed that starting from 2005, and particularly in 2006, the camera shooting of the placed kick highlights has changed considerably, in that the long preparation of the kick (placing the ball, waiting for the placement of the opponents, etc.) is shown more rarely and players' close-ups and medium views are framed instead, resulting into a faster and more dynamic highlight display.

Table 3 - Evolution of the ontology knowledge on the position of the clusters centers for Placed kick highlights 2001-2006

Years	Mean shift	σ^2 shift	Mean cluster radius	n. clusters
2001			3.58	4
2001+2005	3.33	1.34	4.55	6
2001+2005+2006	1.90	1.66	4.44	8
2005			2.86	5
2005+2006	2.89	1.38	4.28	8

In order to test the annotation mechanism for unknown incoming clips we used the subset of automatically detected clips annotated with the 19 concepts, obtained using the perceptual features, described in the previous section.

We selected the shots labeled as *"high angle shots, goal view"* and *"long shots"* for a total amount of 105 shots. This choice was motivated by the fact that these shots are framed using the main camera of the playfield, thus showing the play actions that contain the most significant highlights. We test the automatic high level annotation of these shots using a pictorial ontology containing an initial knowledge related to "Placed Kick", "Shot on Goal" and "Kick off" high level concepts. The set of shots used for the creation of the initial ontology comprised 60 placed kick shots, 90 shot on goal shots and 6 kick off shot. The result of the annotation is shown in Table 4.

Table 4 - High level annotation results

	Total	Correctly Annotated	Miss	False

Placed Kick	12	11	1	12
Shot on Goal	25	20	5	6
Kick off	8	6	2	0

The 80% of shots are correctly annotated. The higher number of false detection of placed kicks is due to source of the video used. In fact, the collection of clips used contains lot of post processed content related to the analysis of the game play. In several cases a video showing an attack action is momentarily frozen and superimposed graphics is used to highlight players' positions and trajectories. The resulting visual features describing these kinds of shots are very similar to the ones of the placed kick since there is the same playfield area framed and the same pattern of camera motion.

The particular video editing used to produce the video collection used (which was a summary of several world cup matches) is responsible also for the miss of the shot on goal. In fact, in many cases only the final part of the attack action leading to a shot on goal is shown resulting in a visual pattern different from the one of the videos commonly broadcasted.

5. CONCLUSIONS AND FUTURE WORK

In this work we described a model for the creation and the maintenance of dynamic perceptual ontologies.

The perceptual part of the ontology has been modeled and created with two complementary approaches. One using low level generic visual features that are able to classify different types of shots, the second using higher level and domain specific features that can refine the results of the shot classification.

As a result of the creation process dynamic pictorial ontologies are enriched with visual prototypes that link the linguistic part of the ontology with the multimedia structural part.

Automatic annotation of video clips using the generated visual concepts has also been presented for soccer domain.

The proposed framework allows the evolution of the pictorial ontology while performing annotation of new videos. We have analyzed the changes of the visual concepts and annotation performances using both low and high level visual features.

Our future work will deal with the extension of multimedia ontologies with the temporal dimension and the consequent temporal evolution and relationships between concepts and entities. In fact, to fully describe the intrinsic nature of video content a knowledge representation has to include formal temporal information, in order to be used for inferring complex concepts based on simple temporal concept relations.

6. REFERENCES

[1] S. Dasiopoulou, V. K. Papastathis, V. Mezaris, I. Kompatsiaris, M. G. Strintzis, "An Ontology Framework For Knowledge-Assisted Semantic Video Analysis and Annotation", *Proc.of 4th International Workshop on SemAnnot at the 3rd International Semantic Web Conference*, Hiroshima, Japan, 2004.

[2] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis M. G. Strintzis, "Knowledge-Assisted Semantic Video Object Detection", *IEEE Transactions on Circuits and Systems for Video Technology*, Special Issue on Analysis and Understanding for Video Adaptation, vol. 15, no. 10, pp. 1210-1224, October 2005

[3] C. Tsinaraki, P. Polydoros, S. Christodoulakis, "Interoperability support for Ontology-based Video Retrieval Applications", *Proc of Conference on Image and Video Retrieval (CIVR)*, 2004

[4] A. Jaimes, J.R. Smith, "Semi-automatic, data-driven construction of multimedia ontologies", *Proc. of International Conference on Multimedia and Expo (ICME)*, 2003.

[5] Th. Athanasiadis, V. Tzouvaras, K. Petridis, F. Precioso, Y. Avrithis Y. Kompatsiaris, "Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content", *Proc. of 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot '05)*, Galway, Ireland, November 2005

[6] "*Towards a common multimedia ontology framework*" analysis of the contribution 27/4/06 report on ACEMedia project

[7] "*Ontological and Epistemological Foundations*", report on DELOS project http://www.idi.ntnu.no/grupper/su/publ/html/totland/ch032.htm

[8] R. Hirschheim, H. Klein, K. Lyytinen, "*Information Systems Development and Data Modeling – Conceptual and Philosophical Foundations*", Cambridge University Press, Cambridge, UK, 289 pages, 1995

[9] E. G. Guba, Y. S. Lincoln, "*Competing Paradigms in Qualitative Research*", in (Denzin and Lincoln, 1994), pp. 105-117, 1994

[10] L. Wittgenstein, "*Tractatus Logico-Philosophicus*", Hypertext Translated from the German by C.K. Ogden http://kfs.org/~jonathan/witt/tlph.html 200

[11] Y. Zhai, J. Liu, X. Cao, A. Basharat, A. Hakeem, S. Ali, M. Shah, C. Grana, R.Cucchiara, "Video Understanding and Content-Based Retrieval", *TREC Video Retrieval Evaluation Online Proceedings*, http://wwwnlpir.nist.gov/projects/tvpubs/tv.pubs.org.html, 2005

[12] C. Grana, R. Cucchiara, "Linear Transition Detection as a Unified Shot Detection Approach" in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, n. 4, pp. 483-489, 2007

[13] M. Bertini, R. Cucchiara, A. Del Bimbo C. Torniai, "Video Annotation with Pictorially Enriched Ontologies". In *Proc. of International Conference on Multimedia and Expo (ICME)*, 2005.

[14] C. Grana, R.Vezzani, R. Cucchiara, "Enhancing HSV Histograms with achromatic points detection in video retrieval", *Proc. of Conference on Image and Video Retrieval (CIVR)*, 2007

[15] D. Zhou, J. Li, H. Zha, "A new Mallows distance based metric for comparing clusterings", *Proc. of the 22nd international conference on Machine Learning*, Bonn, Germany, 2005

[16] M. Bertini, A. Del Bimbo C.Torniai, "Automatic Video Annotation using Ontologies Extended with Visual Information", *Proc. of ACM Multimedia*, November 2005

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

[18] A. Jaimes, B. Tseng, J. Smith, "Modal keywords, Ontologies, and Reasoning for Video Understanding" in *Proc. of Conference on Image and Video Retrieval (CIVR)*, July 2003

[19] A. Benitez, S. – F. Chang, "Automatic Multimedia Knowledge Discovery, Summarization and Evaluation" *IEEE Transactions on Multimedia*, Submitted

[20] J. Kender, M. Naphade, "Visual Concepts for News Story Tracking: Analyzing and Exploiting the NIST TRECVID Video Annotation Experiment" in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 1174-1181, 2005

[21] M. Naphade, J. Smith, J. Tesic, S. Chang, L. Kennedy, A. Hauptmann, J. Curtis, "Large-scale Concepts Ontology for Multimedia", *IEEE Multimedia*, vol. 13, no.3, pp. 86-91, July-Sept 2006

[22] D. Lenat, R. Guha, "*Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*". Reading, MA (USA): Addison-Wesley, 1990.

[23] J. Strintzis, S. Bloehdorn, S. Handschuh, S. Staab, N. Simou, V. Tzouvaras, K. Petridis, I. Kompatsiaris, Y. Avrithis, "Knowledge representation for semantic multimedia content analysis and reasoning," in *Proc. of European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, Nov. 2004.

[24] S. Vembu, M. Kiesel, M. Sintek, S. Bauman, "Towards bridging the semantic gap in multimedia annotation and retrieval," in *Proc. of First International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*, Edinburgh (Scotland), May 2006.

[25] V. Mezaris, I. Kompatsiaris, N. Boulgouris, M. 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.

[26] N. Simou, C. Saathoff, S. Dasiopoulou, E. Spyrou, N. Voisine, V. Tzouvaras, I. Kompatsiaris, Y. Avrithis, S. Staab, "An ontology infrastructure for multimedia reasoning," in *Proc. International Workshop VLBV*, Sardinia (Italy), September, 2005.

[27] D. Kosmopoulos, S. Petridis, I. Pratikakis, V. Gatos, S. Perantonis, V. Karkaletsis, G. Paliouras, "Knowledge Acquisition from Multimedia Content using an Evolution Framework," in *Proc. of 3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI)*, June, 2006.

[28] K. Petridis, S. Bloehdorn, C. Saathoff, N. Simou, S. Dasiopoulou, V. Tzouvaras, S. Handschuh, Y. Avrithis, I. Kompatsiaris, S. Staab, "Knowledge Representation and Semantic Annotation of Multimedia Content," *IEE Proceedings on Vision Image and Signal Processing, Special issue on Knowledge-Based Digital Media Processing*, vol. 153, no. 3, pp. 255-262, June 2006.

[29] C. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Transactions on Multimedia*, vol. 9, no. 5, August, 2007.

[30] C. Grana, D. Bulgarelli, R. Cucchiara, "Video clip clustering for assisted creation of mpeg-7 pictorially enriched ontologies," in *Proc. Second International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Marrakech, Morocco, March 2006.