Semantic Annotation of Images and Videos for Multimedia Analysis

Stephan Bloehdorn[1], Kosmas Petridis[2], Carsten Saathoff[3],
Nikos Simou[4], Vassilis Tzouvaras[4], Yannis Avrithis[4],
Siegfried Handschuh[1], Yiannis Kompatsiaris[2],
Steffen Staab[3], and Michael G. Strintzis[2]

[1] University of Karlsruhe,
Institute AIFB, D-76128 Karlsruhe, Germany
[2] Informatics and Telematics Institute,
GR-57001 Thermi-Thessaloniki, Greece
[3] University of Koblenz-Landau,
Institute for Computer Science, D-56016 Koblenz, Germany
[4] National Technical University of Athens,
School of Electrical and Computer Engineering,
GR-15773 Zographou, Athens, Greece

Abstract. Annotations of multimedia documents typically have been pursued in two different directions. Either previous approaches have focused on low level descriptors, such as *dominant color*, or they have focused on the content dimension and corresponding annotations, such as *person* or *vehicle*. In this paper, we present a software environment to bridge between the two directions. *M-OntoMat-Annotizer* allows for linking low level MPEG-7 visual descriptions to conventional Semantic Web ontologies and annotations. We use *M-OntoMat-Annotizer* in order to construct ontologies that include prototypical instances of high-level domain concepts together with a formal specification of corresponding visual descriptors. Thus, we formalize the interrelationship of high- and low-level multimedia concept descriptions allowing for new kinds of multimedia content analysis and reasoning.

1 Introduction

Representation and semantic annotation of multimedia content have been identified as important steps towards more efficient manipulation and retrieval of visual media. Although new multimedia standards, such as MPEG-4 and MPEG-7 [1], provide important functionalities for the manipulation and transmission of objects and associated metadata, the extraction of semantic descriptions and annotation of the content with the corresponding metadata is out of the scope of these standards and is left to the content manager. This motivates heavy research efforts in the direction of automatic annotation of multimedia content.

Here, we recognize a broad chasm between current multimedia analysis methods and tools on the one hand and semantic annotation methods and tools on the other hand. State-of-the-art multimedia analysis systems are severely limiting themselves by

resorting mostly to visual descriptions at a very low level, e.g. the *dominant color* of a picture. This may be observed even though the need for semantic descriptions that help to bridge the so called *semantic gap* has been acknowledged for a long time [2, 3]. At the same time, the semantic annotation community has only recently started to work into the direction of semantic annotation in the multimedia domain and still remains a long way to go. Work in semantic annotation currently addresses mainly textual resources [4] or simple annotation of photographs [5, 6].

Acknowledging both the relevance of low-level visual descriptions as well as a formal, uniform machine-processable representation [7], we here try to bridge the chasm by providing a semantic annotation framework and corresponding tool, *M-OntoMat Annotizer*, for eliciting and representing knowledge both about the *content domain* and the *visual characteristics* of multimedia data itself. Specifically, MPEG-7 compliant low-level multimedia features are associated with semantic concepts thus forming an a-priori knowledge base.

In the framework we propose, this link between the MPEG-7 visual descriptors and domain concepts is made explicit by means of a conceptualization based on a *prototyping* approach. The core idea of our approach lies in a way to associate concepts with instances that are deemed to be prototypical by their annotators with regard to their visual characteristics. To establish this semantic link we have implemented our framwork in a user-friendly annotation tool, *M-OntoMat-Annotizer*, extending our previous framework for semantic annotations of text [4]. The tool has been built in order to allow content providers to annotate visual descriptors without prior expertise in semantic web technologies or multimedia analysis.

The existence of such a knowledge base may be exploited in a variety of ways. In particular, we envision its exploitation in two modes:

(1) Direct exploitation: In this mode, an application uses the knowledge base directly. For instance, during the semantic annotation process one may gather information like *the blue cotton cloth 4711 in image 12 has a rippled texture described by values 12346546*. Such kind of semantic knowledge may be used later, e.g. for combined retrieval by semantics and similarity in an internet shop. Obviously, such kind of knowledge is expensive to be acquired manually, even when resorting to a user friendly tool. Thus, this kind of knowledge may only be provided for valuable data, such as images or videos of commercial products or of items from museum archives.

(2) Indirect exploitation: In this mode, the a-priori knowledge base 'only' serves as a data set provided to prepare an automatic multimedia analysis tool. For instance, consider the provider of a sports portal offering powerful access to his database on tennis, soccer, etc. He uses semantic annotation of multimedia images or videos in order to prepare an analysis system. For instance, he uses *M-OntoMat-Annotizer* in order to describe the shape and the texture of tennis balls, rackets, nets, or courts and he feeds these descriptions into an analysis system. The system uses the descriptions in order to learn how to tag and relate segments of images and video keyframes with domain ontology concepts. A customer at the portal may then ask the system what it could derive about the images and the videos, e.g. he could ask for all the scenes in which a ball touches a line in a tennis court.

Our long term objectives are dedicated to the indirect exploitation of semantic multimedia annotation as presented in the second paragraph, which is an ongoing comprehensive and complex endeavor, providing a flexible infrastructure for further multimedia content analysis and reasoning, object recognition, metadata generation, indexing and retrieval. In the context of this paper, we only sketch the main steps of our approach.

During image/video analysis, a set of atom-regions is generated by an initial segmentation of images, video sequences and key frames into areas corresponding to salient semantic objects. These objects are also tracked over time while MPEG-7 visual descriptors are extracted for each region. A distance measure between these descriptors and the ones of the prototype instances included in the domain ontology is estimated using a neural network approach. A genetic algorithm then decides the initial labelling of the atom regions with a set of hypotheses, where each hypothesis represents a class from the domain ontology. Finally, a constraint reasoning engine enables the final merging of the regions, while at the same time reducing the number of hypotheses. This approach is generic and applicable to any domain as long as new domain ontologies are designed and made available.

The remainder of the paper is organized as follows: after briefly studying related work in section 2, we present in section 3 an analysis of the initial requirements for the knowledge representation infrastructure both from a knowledge representation and a multimedia analysis point of view. In section 4 we present the general ontology infrastructure design focusing on the multimedia related ontologies and structures. This presentation is complemented by a description of an annotation process needed for initializing the knowledge base with prototype instances of domain concepts in question, including a description of the actual implementation of a user friendly tool to assist this annotation process. Initial results from the knowledge-assisted analysis process, which exploits the developed infrastructure and annotation framework are presented in section 5. We conclude with a summary of our work in section 6.

2   Related Work

In the *multimedia analysis* area, knowledge about multimedia content domains, as for example reported in [8], is a promising approach by which higher level semantics can be incorporated into techniques that capture the semantics through automatic parsing of multimedia content.

Such techniques are turning to knowledge management approaches, including Semantic Web technologies to solve this problem [9]. In [10], semantic entities, in the context of the MPEG-7 standard, are used for knowledge-assisted video analysis and object detection, thus allowing for semantic level indexing. In [11] a framework for learning intermediate level visual descriptions of objects organized in an ontology is presented that aid the system to detect domain objects.

In [12], a-priori knowledge representation models are used as a knowledge base that assists semantic-based classification and clustering. MPEG-7 compliant low-level descriptors are automatically mapped to appropriate intermediate-level descriptors forming a simple vocabulary termed object ontology. Additionally, an object ontology is introduced to facilitate the mapping of low-level to high-level features and allow the

definition of relationships between pieces of multimedia information. This ontology paradigm is coupled with a relevance feedback mechanism to allow for precision in retrieving the desired content.

Work in *semantic annotation* [13] has so far mainly focused on textual resources [4] or simple annotation of photographs [5, 6]. A presentation of an earlier version of *M-OntoMat-Annotizer* can be found in [14].

3   Requirements

The challenge in building a knowledge infrastructure for multimedia analysis and annotation arises from the fact that multimedia data comes in two separate though intertwined layers which need to be appropriately linked. On the one hand, *multimedia layer* deals with the semantics of properties and phenomena related to the presentation of content within the media-data itself, e.g. its spatio-temporal structure or visual features for analysis and is typically hard to understand for people who aren't trained in multimedia analysis. The *content layer*, on the other hand, deals with the semantics of the actual content contained in the media data as it is perceived by the human media consumer. This section analyzes a number of requirements for an integrated knowledge infrastructure and annotation environment for multimedia description, analysis and reasoning. To illustrate some of the requirements, we first present a simple scenario, with focus on direct exploitation:

Multimedia content manager Samantha is working on a project on historic tennis matches. She has to prepare both the metadata infrastructure and the multimedia content. Samantha loads existing general sports ontologies into M-OntoMat-Annotizer and extends them by adding missing concepts of major interest. Next, she points M-OntoMat-Annotizer to images from the project, which are loaded and depicted in the user interface. One after another, Samantha then selects different objects in the images and drags them to the corresponding concepts in the domain ontology. The system extracts visual descriptors for these concepts and stores them in the application memory. Thus, Samantha has used M-OntoMat-Annotizer to describe the tennis domain and to describe the shape and the texture of tennis balls, rackets, nets, or courts.

Note that this simple scenario has focused on simply providing conceptual information *and* the corresponding visual characteristics to the knowledge base which might be exploited directly in the context of the first mode described in section 1.

However, at the same time, the generated data would serve as a valuable a-priori source of information for multimedia analysis tools. These tools would use the descriptions in order to learn how to tag and relate segments of images and video keyframes with the domain ontology concepts in the next step, i.e. in the second mode sketched in section 1, initial results of which are presented in section 5.

3.1   Requirements from Multimedia Analysis

In order to support linking between low level visual information and the higher level content domain, the above example scenario implicitly requires a suitable knowledge infrastructure tailored to multimedia descriptions:

Low-Level Description Representation. In order to represent the visual characteristics associated with a concept, one has to employ several different visual properties, depending on the concept at hand. For instance, in the tennis domain as was described in the scenario, the tennis ball might be described using its shape ("round"), color ("white"), or, in some cases of video sequences, motion. Similarly, a tennis racket has a distinctive and easily recognizable shape.

Support for Multiple Visual Descriptions. Visual Characteristics of domain concepts can not be described using one single instance of the visual descriptors in question. For example, while the net of a tennis racket might be described in terms of its texture only once, its shape heavily depends on the viewing angle and occlusions (e.g. by the player in front of the net). The required conceptualization thus has to provide means for *multiple* prototypical descriptions of a domain concept.

Spatiotemporal Relation Representation. Simple visual properties may be used to model simple concepts. In some cases, however, decomposition of more complex concepts in terms of simpler object parts is desirable. A tennis player, for instance, is difficult to describe using a single shape, motion or texture description; it is more efficient to model and describe the characteristic parts (head, tennis shirt, racket) in terms of their visual properties first, and then define the human player as a spatial configuration of these parts. In other domains like beach holidays, it is more appropriate to describe the entire scene of a picture in terms of its color layout, depicting e.g. the sky at the top, the sand in the middle and the sea at the bottom. In such cases, modelling of spatiotemporal and partonomic relations is required apart from simple visual properties.

Multimedia Structure Representation. The result of the annotation (or content analysis in a next step) should be able to express the structure of a multimedia document itself, depending on the type of document, e.g. image, video, audio, or multimedia presentation. For instance, an image is usually decomposed into a number of still regions corresponding to some semantic objects of interest, while a video clip may be decomposed into shots, each of which into associated moving regions. A hierarchical structure of multimedia segments is thus needed in order to capture all possible types of spatiotemporal or media decompositions and relations.

Alignment with MPEG-7 Standard. The MPEG-7 multimedia content description standard already provides tools for representing fragments of the above information. For instance, the MPEG-7 Visual Part [15] supports color (e.g. dominant colors, color layout), texture, shape (e.g. region/contour-based), and motion (local or global) descriptors. Similarly, the MPEG-7 Multimedia Description Schemes (MDS) [16] supports spatial (directional or topological) and temporal multimedia segment relations, as well as hierarchical structures for multimedia segment decomposition. Given the importance of MPEG-7 in multimedia community, it is evident that in the design of an associated ontology, a large part of its structures should be appropriately captured, aligned and used.

Support for Basic Data Types. Finally, based on the previous requirement, and on the fact that MPEG-7 is built on XML Schema and supported by English-text semantic description but no associated data models, the implementation of an MPEG-7 ontology

using an appropriate formalism like RDF Schema would have to deal with the representation of basic data types like numeric types (integer, float etc.), dates, vectors, arrays and so on. This is a challenging task that is even more important when feature matching algorithms are employed on such data as part of the reasoning process during knowledge-assisted analysis.

3.2   Requirements from Semantic Annotation

The described infrastructure requires appropriate authoring of the domain ontologies with respect to the domain and visual descriptor ontologies.

Associate Visual Features with Concept Descriptions. Visual descriptions are made on the conceptual level, i.e. certain visual descriptors should describe how a certain domain concept is expected to look like. The ontology and annotation framework should model this link in a way that is consistent with current semantic web standards and should avoid 2nd order statements, while

- preserving the ability to use reasoning on the ontology and the knowledge base respectively and
- providing a clear distinction between the visual descriptions of a concept and its instances.

User Friendly Annotation. Domain ontologies are typically edited by trained indexers with little experience in multimedia analysis using standard ontology editing tools. Additionally, maintaining metadata about extracted low level features is cumbersome and error-prone. An annotation framework thus has to integrate:

- management of reference multimedia content (images and videos)
- extraction of suitable low level features for objects depicted in the reference content
- automatic generation of fact statements describing the correspondence between a selected concept and the low level features
- while at the same time hiding the details of these mechanisms to the user behind an easy-to-use user interface.

Modularization. The links between domain ontology concepts and low level feature descriptions should form separate modules of the overall knowledge infrastructure. Specifically, updates of these fact statements should be possible without touching the integrity of the domain ontologies.

Linking into Multimedia. Visual Descriptors contain no information about their location in the original content. This becomes a problem if existing visual descriptors need to be visualized, e.g. to check them for appropriateness or to identify redundant descriptors. Additionally, in order to be able to exploit spatial relationships between objects within multimedia content, the objects have to be linked to the respective regions, they are depicted in. This combines to the more general requirement to provide means to describe regions in terms of their location within the content, i.e. to describe their spatial features, and to link them with objects representing both concepts from the domains and visual descriptors.

Table 1. Matrix of Design Rationales

Requirement / Design Features	Knowledge Infrastructure					M-OntoMat-Annotizer					
Components	Core Ontology	Visual Descriptor Ontology	Multimedia Structure Ontology	Domain Ontologies	Prototyping Approach	Core OntoMat	Annotation Server	Domain Visual Database	Feature Extraction Toolbox	VDE Visual Editor & Media Viewer	VDE Plugin
Ontology Extensions	●			●	●	●	●				
Low-Level Description Representation		●			●				●		●
Support for Multiple Descriptors		●			●		●		●		●
Spatiotemporal Relation Representation	●		●		●		●				
Multimedia Structure Representation			●			●	●	●			
Alignment with MPEG-7 Standards		●	●				●		●		●
Support for Basic Data Types		●							●		
Associate Visual Features with Concept Descriptions					●	●	●			●	●
User Friendly Annotation				●		●		●		●	●
Modularization				●	●	●	●				
Linking into Multimedia								not yet dealt with			

4   Multimedia Annotation Infrastructure Design

Based on the requirements collected in the preceding section, we propose a comprehensive Multimedia Annotation Infrastructure the components of which will be described in this section. Table 1 plots the collected requirements against the infrastructure components.

4.1   Knowledge Infrastructure Design

The requirements presented in the last section point to the challenge that the hybrid nature of multimedia data must be necessarily reflected in the ontology architecture that represents and links both layers. Fig. 1 summarizes the developed knowledge infrastructure[1].
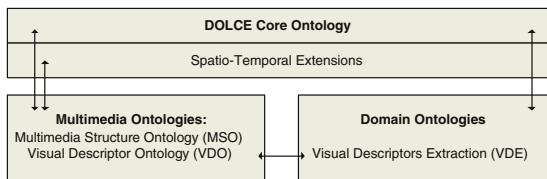

Fig. 1. Ontology Structure Overview

Knowledge Representation Formalisms. Several knowledge representation languages have been developed during the last years as ontology languages in the context of the Semantic Web, each with varying characteristics in terms of their expressiveness, ease of use and computational complexity.

Our framework uses *Resource Description Framework Schema (RDFS)* as modelling language. While RDFS offers sufficient primitives for defining domain models, other parts of the ontology infrastructure either are already encoded in OWL (like DOLCE and the spatio-temporal extensions) or are likely to be leveraged to an appropriate sub-language of OWL at a later stage. This decision also reflects the fact that a full usage of the increased expressiveness of OWL requires specialized and more advanced inference engines, especially when dealing with large numbers of instances with slot fillers, while TBox reasoning is no specific focus of this framework at this point in time.

Core Ontology. The role of the core ontology in this overall framework is to serve as a starting point for the construction of new ontologies, to provide a reference point for comparisons among different ontological approaches and to serve as a bridge between existing ontologies. In our framework, we have used DOLCE [17] for this purpose. DOLCE was explicitly designed as a core ontology, is minimal in that it includes only the most reusable and widely applicable upper-level categories, rigorous in terms of axiomatization and extensively researched and documented.

---

[1] We intend to make these ontologies publicly within 2005.

In a separate module, we have extended the `Region` concept branch of DOLCE to accommodate topological and directional relations between regions of different types, mainly `TimeRegion` and `2DRegion`. Directional spatial relations describe how visual segments are placed and relate to each other in 2D or 3D space (e.g., left and above). Topological spatial relations describe how the spatial boundaries of the segments relate (e.g., touches and overlaps). In a similar way, temporal segment relations are used to represent temporal relationships among segments or events.

Visual Descriptor Ontology. The Visual Descriptor Ontology (VDO) contains the representations of the MPEG-7 visual descriptors, models Concepts and Properties that describe visual characteristics of objects. By the term descriptor we mean a specific representation of a visual feature (color, shape, texture etc) that defines the syntax and the semantics of a specific aspect of the feature. For example, the *dominant color* descriptor specifies among others, the number and value of dominant colors that are present in a region of interest and the percentage of pixels that each associated color value has. Although the construction of the VDO is tightly coupled with the specification of the MPEG-7 Visual Part [18], several modifications were carried out in order to adapt to the XML Schema provided by MPEG-7 to an ontology and the data type representations available in RDF Schema.

The `VDO:VisualDescriptor` concept is the top concept of the Visual Descriptor Ontology and subsumes all modelled visual descriptors. It consists primarily of six subconcepts, one for each category that the MPEG-7 standard specifies. These are: color, shape, texture, motion, localization and basic descriptors. Each of these categories includes a number of relevant descriptors that are correspondingly defined as concepts in the VDO. The only MPEG-7 descriptor category that was modified and does not contain all the MPEG-7 descriptors is the `VDO:BasicDescriptors`.

Multimedia Structure Ontology. The Multimedia Structure Ontology (MSO) models basic multimedia entities from the MPEG-7 Multimedia Description Scheme [16] and mutual relations like decomposition. Within MPEG-7, multimedia content is classified into five types: Image, Video, Audio, Audiovisual and Multimedia. Each of these types has its own segment subclasses. MPEG-7 provides a number of tools for describing the structure of multimedia content in time and space. A number of specialized subclasses are derived from the generic Segment Description Scheme, describing the specific types of multimedia segments, such as video segments, moving regions, still regions and mosaics, which result from spatial, temporal and spatiotemporal segmentation of the different multimedia content types.

Domain Ontologies. In the multimedia annotation framework, the domain ontologies are meant to model the content layer of multimedia content with respect to specific real-world domains, such as sports events like tennis. All domain ontologies are explicitly based on or aligned to the DOLCE core ontology, and thus connected by high-level concepts, what in turn assures interoperability between different domain ontologies at a later stage.

In the context of our work, domain ontologies are created and maintained by content managers or indexers. They are defined in a way to provide a general model of the domain, with focus on the users´ specific point of view. In general, the domain ontology

needs to model the domain in a way, that on the one hand the retrieval of pictures becomes more efficient for a user of a multimedia application and on the other hand the included concepts can also be automatically extracted from the multimedia layer. In other words, the concepts have to be recognizable by automatic analysis methods, but need to remain comprehensible for a human.

Prototype Approach. Describing the characteristics of *concepts* for exploitation in multimedia analysis naturally leads to a meta-concept modeling dilemma. This issue occurs in the sense that using concepts as property values is not directly possible while avoiding 2nd order modelling, i.e. staying within the scope of established standards like OWL DL[2].

In our framework, we propose to enrich the knowledge base with instances of domain concepts that serve as *prototypes* for these concepts. This status is modelled by having these instances also instantiate an additional VDO-EXT:Prototype concept from a separate *Visual Annotation Ontology (VDO-EXT)*. Each of these instances is then linked to the appropriate visual descriptor instances. The approach we have adopted is thus pragmatical, easily extensible and conceptually clean.

4.2   Design of M-OntoMat-Annotizer

While using and referencing the described knowledge representation infrastructure, we have extended the CREAM (CREAting Metadata for the Semantic Web) framework [4] and its reference implementation, OntoMat-Annotizer[3], in order to allow low-level feature annotation. Figure 2 shows the integrated architecture the modules of which are explained in the following in more detail.
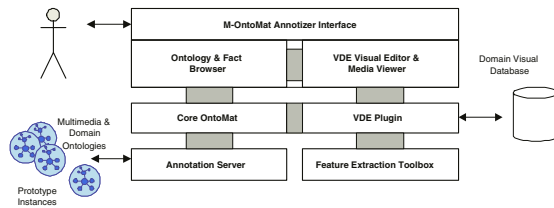
Fig. 2. M-OntoMat-Annotizer and VDE plug-in design architecture

Core OntoMat-Annotizer. OntoMat-Annotizer supports two core applications: (i) it is used as an annotation tool for web pages and (ii) it acts as the basis of an ontology engineering environment. Also, by providing a flexible plug-in interface it offers the pos-

---

[2] The issue of representing concepts as property values is under constant discussion in the Semantic Web Community. As a resource on this topic see Natasha Noy et al.: *Representing Classes As Property Values on the Semantic Web*, W3C Working Draft 21 July 2004, http://www.w3.org/TR/2004/WD-swbp-classes-as-values-20040721/. Note that our approach best resembles approach 2 in this document.

[3] see http://annotation.semanticweb.org/ontomat/

sibility to implement new components and extend the core functionality of OntoMat-Annotizer.

Annotation Server. The annotation server acts in the background and stores the entities of the knowledge base, maintains their mutual references and is responsible for maintaining the overall integrity of the stored entities.

Domain Visual Database. As easy content access is crucial for annotation and content analysis processes, a visual database containing content related to the domain examined and analyzed is always necessary. In aceMedia[4] project, appropriate images and videos are primarily supplied by commercial partners, who actually serve as content providers.

Feature Extraction Toolbox. The actual extraction of the visual descriptors is performed using a feature extraction toolbox, namely the *aceToolbox*, a content preprocessing and feature extraction toolbox developed inside aceMedia project. The ace-Toolbox saves the extracted MPEG-7 Descriptors in XML format.

VDE Visual Editor and Media Viewer. The VDE Visual Editor and Media Viewer presents a graphical interface for loading and processing of visual content (images and videos), visual feature extraction and linking with domain ontology concepts. The interface, as shown in Figure 3, seamlessly integrates with the common OntoMat interfaces. Usually, the user needs to extract the features (multimedia descriptors) of a specific object inside the image/frame. For this reason, the VDE application lets the user draw a region of interest in the image/frame and apply the multimedia descriptors extraction procedure only to the specific selected region. By selecting a specific concept in the OntoMat ontology browser and selecting a region of interest the user can extract and link concepts with appropriate prototype instances by means of the underlying functionalities of the VDE plugin.
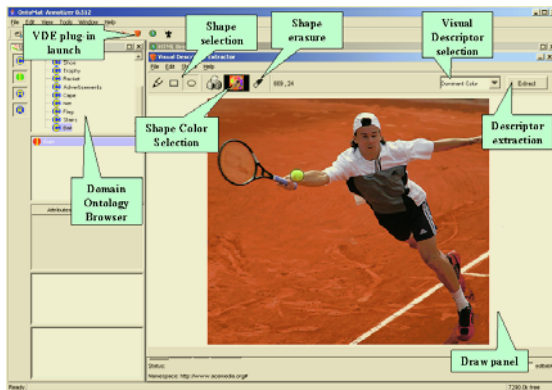

Fig. 3. The VDE plugin into M-OntoMat-Annotizer user interface

---

[4] see http://www.acemedia.org/

VDE Plug-in. The *Visual Descriptor Extraction* (VDE) tool is implemented as a plug-in to OntoMat-Annotizer and is the core component for extending its capabilities and supporting the initialization of ontologies with low-level multimedia features. The VDE plugin manages the overall low-level feature extraction and linking process by communicating with the other components.

4.3   VDE Extraction and Annotation Process

Triggered by the users extraction command, the VDE plugin *extracts* the requested MPEG-7 Descriptors through calls to the Feature Extraction Toolbox. The VDE plug-in supports the *transformation* of the extracted XML multimedia resources into instances of the visual descriptors defined in the VDO, by means of an XSL transformation specification that creates a corresponding descriptor instance for each extraction, which is handed to the knowledge base on the annotation server:

```
<!-- Values of "ScalableColorDescriptor" instance "VDE_INST_0001"-->
<vdo:ScalableColorDescriptor
  rdf:ID="http://www.acemedia.org/ontologies/VDO#VDE_INST_0001">
    <vdo:coefficients> 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 </vdo:coefficients>
    <vdo:numberOfBitPlanesDiscarded> 6 </vdo:numberOfBitPlanesDiscarded>
    <vdo:numberOfCoefficients> 0 </vdo:numberOfCoefficients>
</vdo:ScalableColorDescriptor>
```

Then, the VDE automatically links the newly created visual descriptor instance with the selected domain concept prototype instance on the annotation server:

```
<!-- A prototype instance "Ball_Prototype_1" definition of tennis
concept "Ball". "Ball_Prototype_1" instance "hasDescriptor"
"VDE_INST_0001", which corresponds to the scalable color descriptor
instance above -->
<vdoext:Prototype
  rdf:ID="http://www.acemedia.org/ontologies/VDO-EXT#Ball_Prototype_1">
    <rdf:type rdf:resource="http://www.acemedia.org/ontologies/TENNIS#Ball"/>
    <vdoext:hasDescriptor
      rdf:resource="http://www.acemedia.org/ontologies/VDO#VDE_INST_0001"/>
</vdoext:Prototype>
```

As seen above, these prototype instances are not only instances of the domain concept in question but are also stated to be instances of a separate Prototype concept. The created statements are added to the knowledge base and can be retrieved in a flexible way during analysis. The necessary conceptualizations can be seen as extensions to the VDO (*VDO-EXT ontology*) that link to the core ontology and are implemented in RDF:

```
<!-- Definition of concept "Prototype" as a subclass of Dolce's
"Physical-Object" -->
<rdfs:Class
  rdf:ID="http://www.acemedia.org/ontologies/VDO-EXT#Prototype">
    <rdfs:subClassOf rdf:resource=
      "http://ontology.ip.rm.cnr.it/ontologies/DOLCE-Lite#Physical-Object"/>
</rdfs:Class>

<!-- Definition of relation "hasDescriptor" having Dolce's
"Physical-Object" as domain and VDO's "VisualDescriptor" as range-->
<rdf:Property
  rdf:ID="http://www.acemedia.org/ontologies/VDO-EXT#hasDescriptor ">
    <rdfs:domain rdf:resource=
      "http://ontology.ip.rm.cnr.it/ontologies/DOLCE-Lite#Physical-Object"/>
    <rdfs:range
      rdf:resource="http://www.acemedia.org/ontologies/VDO#VisualDescriptor"/>
</rdf:Property>
```

All the prototype instances can be saved in a RDFS file. The VDE tool saves the domain concept prototype instances together with the corresponding transformed descriptors, *separately* from the ontology file, thus leaving the original domain ontology unmodified.

5   Knowledge-Assisted Analysis

The extracted knowledge base presented above, is playing a central role in automatic semantic multimedia analysis process, through tools currently being developed in ace-Media that automatically analyze content, generate metadata and annotation, and support intelligent content search and retrieval services. Currently, three spatial relations and three low-level descriptors are supported. These include the adjacency (ADJ), below (BEW), and inclusion (INC) relations, and the dominant color (DC), motion (MOV) and compactness (CPS) descriptors.

During preprocessing, color segmentation and motion segmentation are combined to generate a set of over-segmented atom-regions. After preprocessing, assuming for a single image N_R atom regions and a domain ontology of N_O objects, there are $N_R^{N_O}$ possible scene interpretations. A genetic algorithm is used to overcome the computational time constraints of testing all possible configurations [19].

The degree of matching between regions, in terms of low-level visual and spatial features respectively, is defined in an interpretation function used for the genetic algorithm fitness function and is based on a back-propagation neural network. When the task is to compare two regions based on a single descriptor, several distance functions can be used; however, there is not a single one to include all descriptors with different weight on each. This is a problem that is handled by the neural network. Its input consists of the low-level descriptions of both an atom region and an object model, while its
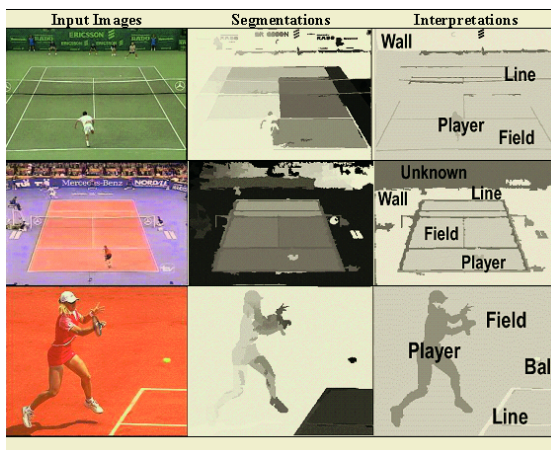

Fig. 4. Tennis domain results

response is the estimated normalized distance between the atom region and the model. A training set is constructed using the descriptors of a set of manually labelled atom regions and the descriptors of the corresponding object models. Figure 4 illustrates example results from the sports domain, where the system output is a segmentation mask outlining the semantic description of the scene.

6   Summary

In this paper, an integrated infrastructure for semantic multimedia content annotation was presented. This framework comprises ontologies for the description of low-level audio-visual features and for linking these descriptions to concepts in domain ontologies based on a prototype approach. This prototype approach avoids the well-known problems introduced by Meta-Concept Modelling, and thus preserves the ability to use OWL DL compliant reasoning techniques on the annotation meta-data.

The generation of the visual descriptors and the linking with the domain concepts is embedded in an user-friendly tool, which hides analysis-specific details from the user. Thus, the definition of appropriate visual descriptors can be accomplished by domain experts, without the need to have a deeper understanding of ontologies or low-level multimedia representations. In allowing annotation and linking of concept prototype instances with more than one extracted descriptors, the system is flexible with respect to analysis requirements. In allowing multiple prototypical instantiations of the a concept, the system is flexible with respect to varying visual characteristics of objects.

An important issue in the actual annotation procedure, is the selection of appropriate descriptors for extraction, valuable for the further analysis process. Depending on the results, the knowledge-assisted analysis process adjusts its needs and guides the extraction procedure, providing constant feedback on the concepts that have to be populated, how many prototype instances are necessary for each concept, which descriptors are helpful for the analysis of a specific concept etc.

Finally, despite the early stage of multimedia analysis experiments, first results based on the ontologies presented in this work are promising and show that it is possible to apply the same analysis algorithms to process different kinds of images or video, by simply employing different domain ontologies. Apart from visual descriptions and relations, future focus will concentrate on the creation of rules to assist reasoning in order to detect more complex events. The examination of the interactive process between ontology evolution and use of ontologies for content analysis will be the target of our future work, in the direction of handling the semantic gap in multimedia content interpretation.

References

1. S.-F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6):688–695, June 2001.
2. O. Mich R. Brunelli and C.M. Modena. A survey on video indexing. *Journal of Visual Communications and Image Representation*, 10:78–112, 1999.
3. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12).
4. Siegfried Handschuh and Steffen Staab. Cream - creating metadata for the semantic web. *Computer Networks*, 42:579–598, AUG 2003. Elsevier.
5. J. Wielemaker A.Th. Schreiber, B. Dubbeldam and B.J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, May/June 2001.
6. L. Hollink, A.Th. Schreiber, J. Wielemaker, and B. Wielinga. Semantic annotation of image collections. In *Proceedings of the K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation, Florida*, 2003.
7. P. Wittenburg D. Thierry and H. Cunningham. The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment. In *Proc. ACL/EACL Workshop on Human Language Technology and Knowledge Management*, Toulouse, France, 2001.
8. J. Hunter, J. Drennan, and S. Little. Realizing the hydrogen economy through semantic web technologies. *IEEE Intelligent Systems Journal - Special Issue on eScience*, 19:40–47, 2004.
9. A. Yoshitaka, S. Kishida, M. Hirakawa, and T. Ichikawa. Knowledge-assisted content-based retrieval for multimedia databases. *IEEE Multimedia*, 1(4):12–21, Winter 1994.
10. R. Tansley, C. Bird, W. Hall, P. Lewis, and M. Weal. Automating the linking of content and concept. In *Proc. ACM Int. Multimedia Conf. and Exhibition (ACM MM-2000)*, Oct./Nov. 2000.
11. Nicolas Maillot, Monique Thonnat, and Céline Hudelot. Ontology based object learning and recognition: Application to image retrieval. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004), 15-17 November 2004, Boca Raton, FL, USA*, pages 620–625, 2004.
12. I. Kompatsiaris, V. Mezaris, and M. G. Strintzis. *Multimedia content indexing and retrieval using an object ontology*. Multimedia Content and Semantic Web - Methods, Standards and Tools, Editor G.Stamou, Wiley, New York, NY, 2004.
13. Siegfried Handschuh and Steffen Staab, editors. *Annotation for the Semantic Web*. IOS Press, 2003.
14. Stephan Bloehdorn, Steffen Staab, Siegfried Handschuh, Yannis Avrithis, Vasilis Tzouvaras, Nikos Simou, Michael G. Strintzis, Yiannis Kompatsiaris, and Kosmas Petridis. Knowledge representation for semantic multimedia content analysis and reasoning. In *Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT)*, NOV 2004.
15. T. Sikora. The MPEG-7 Visual standard for content description - an overview. *IEEE Trans. on Circuits and Systems for Video Technology, special issue on MPEG-7*, 11(6):696–702, June 2001.
16. ISO/IEC 15938-5 FCD Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Scemes, March 2001, Singapore.

17. A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening Ontologies with DOLCE. In *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Proceedings of the 13th International Conference on Knowledge Acquisition, Modeling and Management, EKAW 2002*, volume 2473 of *Lecture Notes in Computer Science*, Siguenza, Spain, 2002.

18. ISO/IEC 15938-3 FCD Information Technology - Multimedia Content Description Interface - Part 3: Visual, March 2001, Singapore.

19. N. Voisine, S. Dasiopoulou, V. Mezaris, E. Spyrou, T. Athanasiadis, I. Kompatsiaris, Y. Avrithis, and M.G. Strintzis. Knowledge-Assisted Video Analysis Using A Genetic Algorithm. In *Proc. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005)*, April 2005.