Knowledge Representation and Semantic Annotation of Multimedia Content

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

[1] Informatics and Telematics Institute, GR-57001 Thermi-Thessaloniki, Greece
[2] University of Karlsruhe, Institute AIFB, D-76128 Karlsruhe, Germany
[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. Knowledge representation and annotation of multimedia documents typically have been pursued in two different directions. Previous approaches have focused either on low level descriptors, such as *dominant color*, or on the semantic content dimension and corresponding manual annotations, such as *person* or *vehicle*. In this paper, we present a knowledge infrastructure and a experimentation platform for semantic annotation to bridge the two directions. Ontologies are being extended and enriched to include low-level audiovisual features and descriptors. Additionally, we present a tool that allows for linking low-level MPEG-7 visual descriptions to ontologies and annotations. This way we construct ontologies that include prototypical instances of high-level domain concepts together with a formal specification of the corresponding visual descriptors. This infrastructure is exploited by a knowledge-assisted analysis framework that may handle problems like segmentation, tracking, feature extraction and matching in order to classify scenes, identify and label objects, thus automatically create the associated semantic metadata.

1 Introduction

Representation and semantic annotation of multimedia content have been identified as important steps towards more efficient manipulation and retrieval of visual media. Today, new multimedia standards such as MPEG-4 [1] and MPEG-7 [2], provide important functionalities for manipulation and transmission of objects and associated metadata. The extraction of semantic descriptions and annotation of the content with the corresponding metadata though, is out of the scope of these standards and is still left to the content manager. This motivates heavy research efforts in the direction of automatic annotation of multimedia content [3].

A gap is acknowledged between existing multimedia analysis methods and tools on one hand and semantic description, annotation methods and tools on the other. The state-of-the-art multimedia analysis systems are severely limited to resorting mostly to visual descriptions at a very low level, e.g. the *dominant color* of a picture [4]. However, ontologies that express key entities and relationships of multimedia content in a formal machine-processable representation can help bridging the *semantic gap* [5, 6] between the automatically extracted low-level arithmetic features and the high-level human understandable semantic concepts. At the same time, the semantic annotation community has only recently started working into the direction of tackling the many problems of semantic annotation in the multimedia domain [7, 8].

Acknowledging the relevance between low-level visual descriptions and formal, uniform machine-processable representations [9], we try to bridge this gap by providing a knowledge infrastructure design focusing on the multimedia related ontologies and structures. Additionally, we present a prototypical annotation framework and corresponding tool, *M-OntoMat Annotizer* [10] that is capable of eliciting and representing knowledge both about the *content domain* and the *visual characteristics* of multimedia data itself. More specifically, MPEG-7 visual descriptors, are associated to semantic concepts thus forming an a-priori knowledge base.

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, but requires manual intervention. 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, which is presented in detail in this paper, the a-priori knowledge base serves as a data set provided to set up and train an automatic multimedia analysis framework. For instance, consider the providers of a sports portal offering powerful access to their database on tennis, soccer etc. They use the ontology infrastructure and the prototypical annotation of multimedia images or videos in order to configure an analysis system. More specifically, they use M-OntoMat-Annotizer to describe the color, shape and texture of tennis balls, rackets, nets or courts and they feed these descriptions into the system. The system can then use the descriptions to learn how to automatically tag and relate segments of more images and video key-frames in the database with domain ontology concepts. Customers at the portal may then ask the system to retrieve specific objects or events in images and videos, e.g. they could ask for *all the scenes in which a ball touches a line in a tennis court*.

More specifically, our analysis system includes methods that automatically segment images, video sequences and key frames into areas corresponding to salient semantic objects (e.g. cars, road, people, field etc), track these objects over time, and provide a flexible infrastructure for further analysis of their relative motion and interactions, as well as object recognition, metadata generation, indexing and retrieval. Recognition is then performed by comparing existing prototypical descriptions to lower-level features extracted from the signal (image/video), thus identifying objects and their relations in the multimedia content.

During image/video analysis, a set of atom-regions is generated by an initial segmentation while visual descriptors and spatial relations 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 for distance weighting. Finally, a genetic algorithm decides the labeling of the atom regions with a set of hypotheses, where each hypothesis represents a concept from the domain ontology. 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, section 3 presents an analysis of the initial requirements of the knowledge 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 M-OntoMat-Annotizer that initializes the knowledge base with prototypical descriptor instances of domain concepts. The knowledge-assisted analysis platform, which is exploiting the developed infrastructure, and a comprehensive evaluation 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 [11], is a promising approach by which higher level semantics can be incorporated into techniques that capture them through automatic parsing of multimedia content. In order to solve this problem, such techniques are turning to knowledge management approaches, including Semantic Web technologies. In [12], ontology-based semantic descriptions of images are generated based on appropriately defined rules that associate MPEG-7 low-level features to the concepts included in the ontologies. The architecture presented in [13] consists of an audio-visual ontology in compliance with the MPEG-7 specifications and corresponding domain ontologies. A semantic repository is used for storing and querying the ontologies and the statements, while the use of a reasoner enables enhanced inference services.

In [14], 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 [15], a framework for learning intermediate level visual descriptions of objects organized in an ontology is presented that aids the system to detect domain objects. In [16], 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* [17] has so far mainly focused on textual resources [18] or simple annotation of photographs [8, 19].

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, the *multimedia layer* deals with the semantics of properties and phenomena related to the presentation of content within the media-data itself, e.g. its spatiotemporal structure or visual features for analysis and is typically hard to understand for people who aren't trained in multimedia analysis. On the other hand, the *content layer* deals with the semantics of the actual content contained in the media data as it is perceived by the human media consumer. This section describes a number of requirements for an integrated knowledge infrastructure and annotation environment for multimedia description and analysis.

3.1 Requirements from Multimedia Analysis

Supporting the linking between low-level visual information and the higher level content domain, 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, the tennis ball might be described using its shape (e.g. round), color (e.g. white), or in case of video sequences, its motion.

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. 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 domains like beach holidays however, 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 sea in the middle and the sand at the bottom. In such cases, modeling of spatiotemporal and patronomic relations is required apart from simple visual properties.

Multimedia Structure Representation The result of the content analysis should be able to express the structure of a multimedia document itself, depending on the type of document. 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 the 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* [20] supports color, texture, shape and motion descriptors. Similarly, the *MPEG-7 Multimedia Description Schemes (MDS)* [21] supports spatial 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 MPEG-7 should be appropriately captured, aligned and used.

3.2 Requirements from Knowledge Representation

The described infrastructure requires appropriate authoring of the domain ontologies with respect to the corresponding multimedia ontologies.

Associate visual features with concept descriptions Visual descriptions are made on the conceptual level, i.e. certain visual descriptors should describe how a specific domain concept is expected to look like. The ontology and prototypical annotation framework should model this link in a way that is consistent with current semantic web standards, while preserving the ability to use reasoning on the ontologies 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 multimedia content, 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 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 visual descriptors should be possible without touching the integrity of the domain ontologies.

4 Knowledge Representation

Based on the requirements collected above, we propose a comprehensive ontology infrastructure, the components of which will be described in this section. These requirements point to the challenge that the hybrid nature of multimedia data must be necessarily reflected in the ontology architecture that represents and links both the multimedia and the content layer. Fig. 1 summarizes the developed knowledge infrastructure.

4.1 Ontology Infrastructure

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 modeling language. This decision reflects the fact that a full usage of the increased expressiveness of *Web Ontology Language (OWL)* requires specialized and more advanced inference engines that are still not in mature state, especially when dealing with large numbers of instances with slot fillers.

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* [22] for this purpose. DOLCE is 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.

In a separate module, we have carefully 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 visual descriptors, models concepts and properties that describe visual characteristics of objects. Although the construction of the VDO is tightly coupled with the specification of the MPEG-7 Visual Part [20], several modifications were carried out so that VDO could adapt the XML Schema specification provided by MPEG-7 to the data type representations available in RDF Schema.

The `VDO:VisualDescriptor` concept is the top concept of the VDO and subsumes all modeled 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*. As an example, Fig. 2 illustrates the `ColorDescriptor` branch of the VDO. 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 [21] 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. These subclasses describe 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. MPEG-7 provides a number of tools for describing the structure of multimedia content in time and space. Multimedia resources can be segmented or decomposed into sub-segments through four types of decomposition: *spatial, temporal, spatiotemporal* and *media source*.

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 general, domain ontologies need to model the domain in a way that on the one hand the retrieval of content 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 2^{nd} order modeling, i.e. staying within the scope of OWL DL[1].

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 modeled by having these instances also instantiate an additional `Prototype` concept from a separate *Visual Annotation Ontology (VDO-EXT)*. Each of these instances is then linked to the appropriate

[1] The issue of representing concepts as property values is under constant discussion in the Semantic Web Community. As a resource on this topic see [23]. Note that our approach best resembles approach 2 in this document.

visual descriptor instances. The approach we have adopted is thus pragmatical, easily extensible and conceptually clean.

4.2 M-OntoMat-Annotizer framework

In order to exploit the ontology infrastructure presented above and enrich the domain ontologies with multimedia descriptors the usage of a tool is necessary. The implemented framework is called *M-OntoMat-Annotizer*[2] (M stands for Multimedia) [10]. The development was based on an extension of the *CREAM (CREAting Metadata for the Semantic Web)* framework [18] and its reference implementation *OntoMat-Annotizer*[3].

For this reason, the *Visual Descriptor Extraction Tool (VDE)* tool was 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 plug-in manages the overall low-level feature extraction and linking process by communicating with the other OntoMat-Annotizer components.

The VDE visual editor and media viewer presents a graphical interface for loading and processing of visual content, visual features extraction and linking with domain ontology concepts. The interface, as shown in Fig. 3, seamlessly integrates with the common OntoMat-Annotizer ones. Usually, the user needs to extract the visual features (i.e. descriptors included in the VDO) of a specific object inside the image/frame. M-OntoMat-Annotizer lets the user draw a region of interest in the image/frame and apply the multimedia descriptor extraction procedure only to the specific selected region. By specifying an instance of a concept in the ontology browser and selecting a region of interest the user can extract and link appropriate visual descriptor instances with instances of domain concepts that serve as *prototypes* for these concepts.

[2] see `http://www.acemedia.org/aceMedia/results/software/m-ontomat-annotizer.html`

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

As discussed earlier, these prototypes 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 RDFS. M-OntoMat-Annotizer saves the domain concept prototype instances together with the corresponding descriptors, in a separate RDFS file and leaves the original domain ontology unmodified.

5 Knowledge-Assisted Multimedia Analysis

The knowledge base described above serves as the primary reference resource for the analysis process, which in turn leads to the semantic annotation of the examined multimedia content.

5.1 Platform Architecture

The general architecture scheme of our knowledge-assisted analysis platform is given in Fig. 4. The core of the architecture is defined by the region adjacency graph that holds the region-based representation of the input content during the analysis process. Each node of the graph corresponds to an atom-region while a graph edge represents the link between two regions, holding the overall neighboring information. Four visual descriptors are supported and could be extracted for each region: *dominant color* **(**DC**)**, *region shape* **(**RS**)**, *motion* **(**MOV**)** and *compactness* **(**CPS**)**.

For the analysis purposes, the system needs to have full access to the overall knowledge base consisting of the domain concept prototype instances together with the associated visual descriptors. These instances are applied as references to the analysis algorithms and are compared to the corresponding descriptors of each node of the graph. Then the platform extracts the semantic concepts that are linked to the regions of the image or video shot.

For the actual retrieval of prototypes and its descriptor instances, the *OntoBroker* [4] engine is used and deals with the necessary queries to the knowledge base, since it supports loading of RDFS ontologies. OntoBroker needs to load the domain ontologies where the high-level concepts are defined, the VDO where the low-level visual descriptors are present, and the prototype instance files that include the actual knowledge base and provide the linking of domain concepts with descriptor instances. Appropriate queries are defined succeeding the retrieval of specific values from the stored descriptors and concepts. The OntoBroker's query language is *F-Logic* [5]. F-Logic is both a representation language that can model ontologies and a query language that can be used to query OntoBroker's loaded knowledge.

The analysis process starts with a preprocessing step where region-based segmentation [24] and motion segmentation [25] are combined. Region motion estimation is performed using both the microblock motion vectors extracted from the compressed domain and the global motion compensation. The latter is necessary when the camera is moving and is based on estimating the eight parameters of the bilinear motion model for camera motion, using an iterative rejection procedure [26]. The compactness descriptor is calculated by the area and the perimeter of the region. Dominant color and region shape are extracted for each region as well as spatial relations (such as `above`, `below`, `is-included-in`) between adjacent regions.

After preprocessing, assuming that for a single image N_R atom regions are generated and there is 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 con-

[4] see `http://www.ontoprise.de/products/ontobroker_en`

[5] see `http://www.ontoprise.de/documents/tutorial_flogic.pdf`

figurations [27]. In this approach, each individual represents a possible interpretation of the examined scene, i.e the identification of all atom regions. In order to reduce the search space, the initial population is generated by allowing each gene to associate the corresponding atom-region only with those objects that the particular atom-region is most likely to represent.

The following functions are defined to estimate the degree of matching in terms of low-level visual and spatial features respectively between an atom-region r_i and an object concept o_j.

- the *matching interpretation function* $\mathcal{I}_M^t(r_i, o_j)$, assuming that gene g_t associates region r_i with object o_j, to provide an estimation of the degree of matching between o_j and r_i. $\mathcal{I}_M^t(r_i, o_j)$ is calculated using appropriate descriptor distance functions and is subsequently normalized so that $\mathcal{I}_M^t(r_i, o_j)$ belongs to $[0, 1]$, with a value of 1 indicating a perfect match.
- the *relation interpretation function* $\mathcal{I}_R^t(r_i, r_k)$, which provides an estimation of the degree to which the spatial relation between atom-regions r_i and r_k satisfies the relation $\mathcal{R}$ defined in the ontology between the objects to which r_i and r_k are respectively mapped to by gene g_t.

Since each individual represents the scene interpretation, the fitness function has to consider the above-defined low-level visual and spatial matching estimations for all atom-regions. As a consequence the employed *fitness function* is defined as:

$$Fitness(g_t) = (\sum_{i}^{N_R} \mathcal{I}_M^t(r_i, o_m)) \prod_{i}^{N_R} \prod_{j \in S_i} \mathcal{I}_R^t(r_i, r_j)$$

where S_i denotes the set of neighboring atom-regions of r_i, since the spatial relations used have been defined only for regions with connected boundaries. It follows from the above definitions that the optimal solution is the one that maximizes the fitness function. In our implementation, the *roulette wheel selection* genetic operator was used, in which individuals are given a probability of being selected that is directly proportional to their fitness and uniform crossover, where genes of the parent chromosomes are randomly copied.

Our approach to implement the matching interpretation function $\mathcal{I}_M^t$ used for the fitness function, 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 the neural network handles. Its input consists of the low-level descriptions of both an atom region and an object prototype, while its response is the estimated normalized distance between the atom region and the prototype. A training set is constructed using the descriptors of a set of manually-labeled atom regions and the descriptors of the corresponding object prototypes. The network is trained under the assumption that all descriptors are equally important. Moreover, the distance of an atom region that belongs to the training set is minimum for the associated prototypes and maximum for all others. This distance is then used for the interpretation function $\mathcal{I}_M^t$.

Following the above labeling procedure, each region is assigned to a set of plausible hypotheses, i.e. a set of concepts that combine lowest distance from the corresponding prototype descriptor instances and spatial relations consistency. To reach the final semantic description, the interpretation with the highest fitness score is used for labeling the image regions.

5.2 Experimental Results and Evaluation

The generated metadata express the structure and semantics of the analyzed content, ie. a number of still regions or shots accompanied by a semantic label. They are produced in RDFS format following the definitions of the MSO, thus they are tightly coupled with the integrated knowledge infrastructure. As illustrated in Fig. 5 and 6, the system output also includes a segmentation mask outlining the semantic description of the scene. The different colors assigned to the generated atom-regions correspond to the object classes defined in the domain ontology.

Concerning the evaluation process, a simple approach was selected. The examined image is partitioned into a fixed number of blocks and is compared against the respective ground truth annotation. In the current implementation, the grid size is dynamically calculated for each examined image, so as to result in partitioning the image into 64 blocks (8*8 grid). Consequently, the obtained evaluation is rather detailed and adequately indicative of the achieved annotations accuracy.

As performance measures, *precision* and *recall* from the *information retrieval (IR)* field are used. Precision specifies the percentage of correctly retrieved concepts over all retrieved concepts, whereas recall specifies the percentage of correctly retrieved concepts over all correct concepts. In other words, precision indicates how many incorrect concepts were retrieved, and recall how many correct concepts were missed. Following the framework above, ground truth is created based on a 8*8 grid. Each block of the grid is then mapped to the regions it overlaps and the block labels mapped to a region that contains them are counted.

Subsequently, both precision and recall are calculated. A known problem of precision and recall is that although they have a strong relationship, they are two distinct measures. Therefore *F-Measure* is also computed denoting the harmonic mean of precision (p) and recall (r), i.e. $\mathcal{F} = 2pr/(p + r)$, where in contrast to the arithmetic mean, it only gets large if both precision and recall become large.

For the evaluation of the knowledge-assisted analysis platform, a set of 200 images belonging to the *holiday-beach* domain was selected, while the *Sky, Sea, Sand* and *Person* concepts were examined. The evaluation results are illustrated in Table 1.

6 Conclusion and Future Work

In this paper, an integrated infrastructure for semantic annotation of multimedia content was presented. This framework comprises ontologies for the description of low-level visual features and for linking these descriptions to concepts in domain ontologies based on a prototype approach. This approach avoids the well-known problems introduced by meta-concept modeling, and thus preserves the ability to use OWL DL compliant reasoning techniques on the annotation metadata. The generation of the visual descriptors and the linking with domain concepts is embedded in a user-friendly tool that 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.

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. Despite the early stage in 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.

Due to the fact that it is not possible to decide on the correct label solely based on low-level features, a post-processing is required to incorporate further knowledge into the labeling process. Furthermore, due to segmentation numerical limitations, i.e. illumination variations, objects with non-homogeneous parts etc, semantically meaningful regions might be segmented into a number of smaller ones. Such over-segmented atom-regions need to be merged in order to achieve a segmentation and consequently a labeling that corresponds to semantically meaningful regions.

Hence, a future direction in our implementation will be to consider introducing reasoning mechanisms and using additional spatiotemporal context knowledge to allow for further processing. Reasoning will be used to refine the hypotheses sets, i.e. exclude labels that do not fit into the spatiotemporal context of related segments, and to merge segments that belong to the same semantic entity, e.g. by exploiting topological knowledge, partonomic relations etc. Additionally, appropriate rules will be defined to drive the reasoning process in order to detect more complex events and support improved semantic description decisions.

Finally, the examination of the interactive process between ontology evolution and use of ontologies for content analysis will also be the target of our future work, in the direction of handling the semantic gap in multimedia content interpretation.

Acknowledgements This research was partially supported by the European Commission under contract FP6-001765 aceMedia. The expressed content is the view of the authors but not necessarily the view of the aceMedia project as a whole.

References

1. Overview of the MPEG-4 Standard. Technical report, ISO/IEC JTC1/SC29/WG11 N1730, Stockholm Jul. 1997.
2. 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.
3. S.-F. Chang. The holy grail of content-based media analysis. *IEEE Multimedia*, 9(2):6–10, Apr.-Jun. 2002.
4. A. Yoshitaka and T. Ichikawa. A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):81–93, Jan/Feb 1999.
5. O. Mich R. Brunelli and C.M. Modena. A survey on video indexing. *Journal of Visual Communications and Image Representation*, 10:78–112, 1999.

6. 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).
7. Laura Hollink, Giang Nguyen, Guus Schreiber, Jan Wielemaker, Bob Wielinga, and Marcel Worring. Adding Spatial Semantics to Image Annotations. In *Proceedings of the 4th International Workshop on Knowledge Markup and Semantic Annotation at 3rd International Semantic Web Conference*, November 2004.
8. J. Wielemaker A.Th. Schreiber, B. Dubbeldam and B.J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, May/June 2001.
9. T. Declerck, P. Wittenburg, and H. Cunningham. The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment. In *Proceedings of the ACL/EACL Workshop on Human Language Technology and Knowledge Management*, Toulouse, France, 2001.
10. S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, I. Kompatsiaris, S. Staab, and M.G. Strintzis. Semantic Annotation of Images and Videos for Multimedia Analysis. In *Proceedings of the 2nd European Semantic Web Conference (ESWC 2005)*, May 2005.
11. 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.
12. 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.
13. R. Troncy. Integrating Structure and Semantics into Audio-Visual Documents. In *Proceedings of the 2nd International Semantic Web Conference (ISWC 2003)*, October 2003.
14. R. Tansley, C. Bird, W. Hall, P. Lewis, and M. Weal. Automating the linking of content and concept. In *Proceedings of the ACM Int. Multimedia Conf. and Exhibition (ACM MM-2000)*, Oct./Nov. 2000.
15. 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.
16. 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.
17. Siegfried Handschuh and Steffen Staab, editors. *Annotation for the Semantic Web*. IOS Press, 2003.
18. Siegfried Handschuh and Steffen Staab. Cream - creating metadata for the semantic web. *Computer Networks*, 42:579–598, AUG 2003. Elsevier.
19. 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.
20. ISO/IEC 15938-3 FCD Information Technology - Multimedia Content Description Interface - Part 3: Visual, March 2001, Singapore.
21. ISO/IEC 15938-5 FCD Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Scemes, March 2001, Singapore.
22. 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.

23. 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/).
24. T. Adamek, N.O'Connor, and N.Murphy. Region-based Segmentation of Images Using Syntactic Visual Features. In *Proc. Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2005*, Montreux, Switzerland, April 13-15 2005.
25. V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, and M.G. Strintzis. Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(5):606–621, May 2004.
26. T. Yu and Y. Zhang. Retrieval of video clips using global motion information. *Electronics Letters*, 37(14):893–895, July 2001.
27. N. Voisine, S. Dasiopoulou, F. Precioso, V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. A Genetic Algorithm-based Approach to Knowledge-assisted Video Analysis. In *Proceedings of the IEEE International Conference on Image Processing (ICIP 2005)*, September 2005.

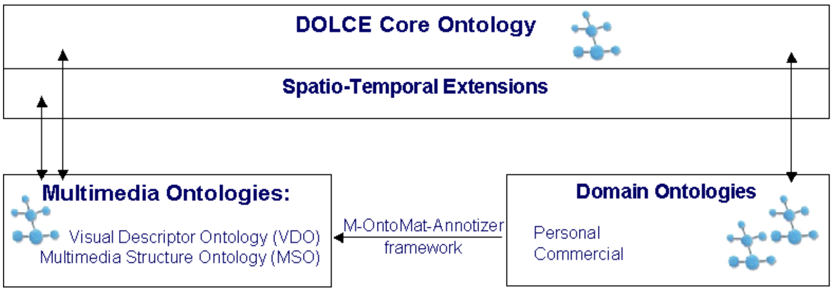

Fig. 1. Ontology Structure Overview

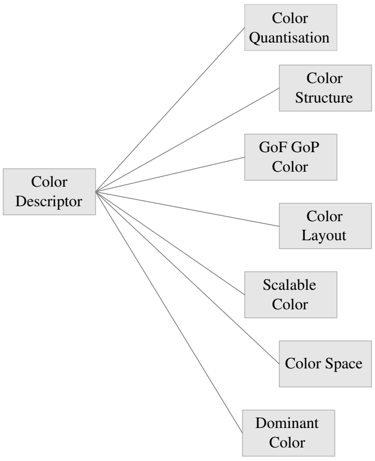

Fig. 2. The `VDO:ColorDescriptor` hierarchy

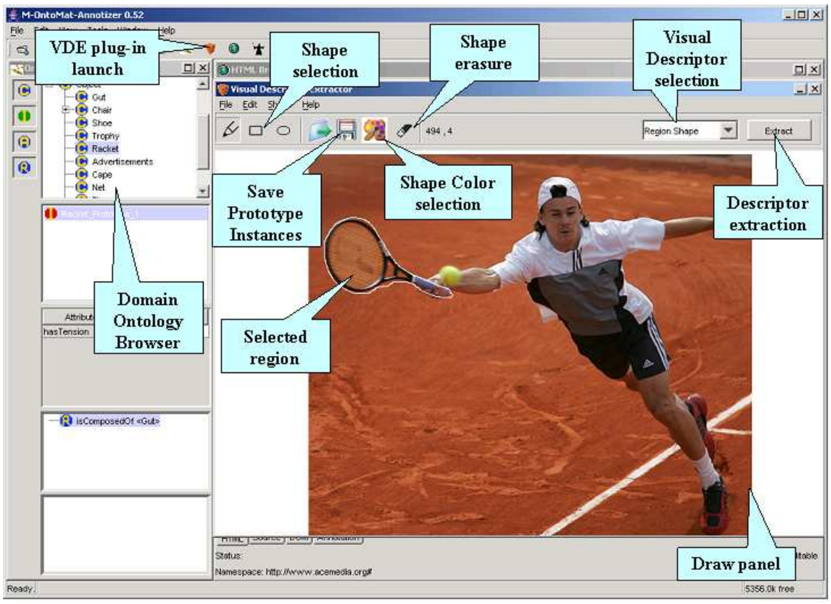

Fig. 3. The M-OntoMat-Annotizer user interface

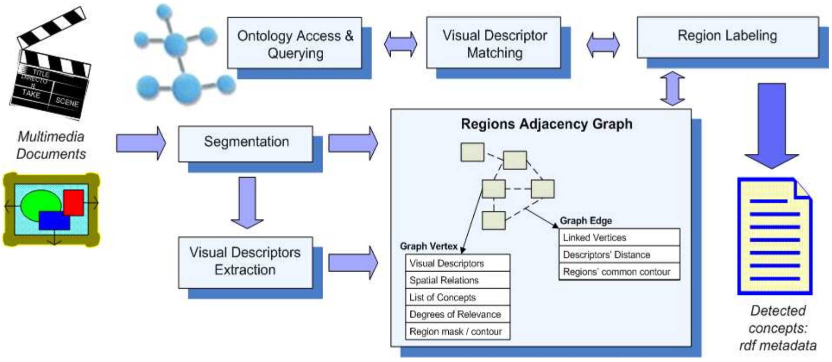

Fig. 4. Knowledge-assisted analysis architecture

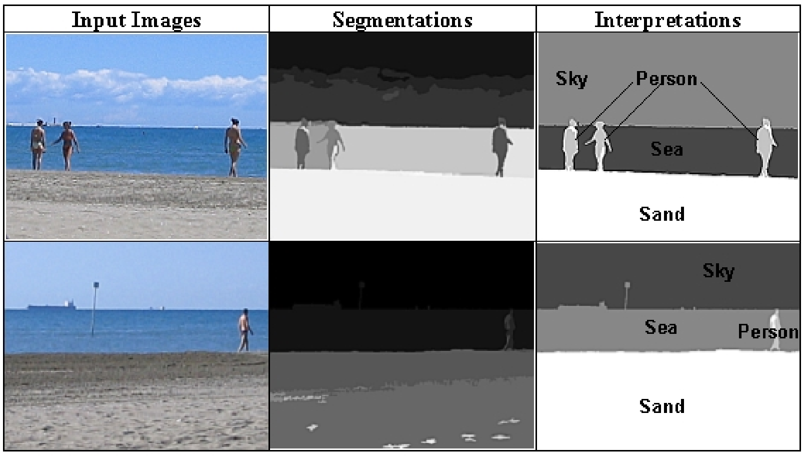

Fig. 5. Holiday-Beach domain results

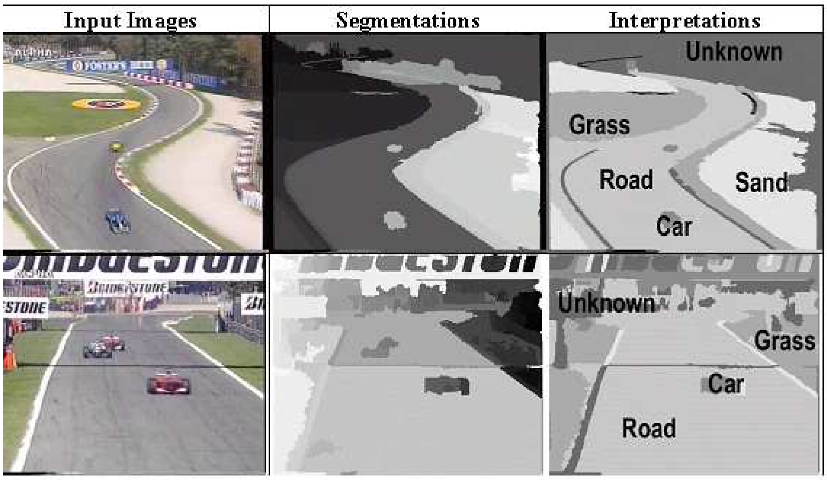

Fig. 6. Formula One domain results

Concept	Precision	Recall	F
Sky	0.96	0.93	0.94
Sea	0.94	0.85	0.89
Sand	0.88	0.78	0.83
Person	0.88	0.63	0.73

Table 1. Knowledge-assisted analysis evaluation results