Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content

Thanos Athanasiadis[1], Vassilis Tzouvaras[1], Kosmas Petridis[2], Frederic Precioso[2],
Yannis Avrithis[1] and Yiannis Kompatsiaris[2]

[1] National Technical University of Athens, School of Electrical and Computer Engineering,
GR-15773 Zographou, Athens, Greece
[2] Informatics and Telematics Institute, GR-57001 Thermi-Thessaloniki, Greece

Abstract. In this paper we discuss the use of knowledge for the automatic extraction of semantic metadata from multimedia content. For the representation of knowledge we extended and enriched current general-purpose ontologies to include low-level visual features. More specifically, we implemented a tool that links MPEG-7 visual descriptors to high-level, domain-specific concepts. For the exploitation of this knowledge infrastructure we developed an experimentation platform, that allows us to analyze multimedia content and automatically create the associated semantic metadata, as well as to test, validate and refine the ontologies built. We pursued a tight and functional integration of the knowledge base and the analysis modules putting them in a loop of constant interaction instead of being the one just a pre- or post-processing step of the other.

1 Introduction

Recent advances in computing technologies have made available vast amount of digital video content, resulting in a growing research interest in the extraction of semantic metadata, that can provide a description in a conceptual level. At the same time, the fundamental prerequisite of the Semantic Web is *making content machine-understandable*; however the semantic annotation community has only recently worked towards this direction [1, 2]. Although significant progress has been made on automatic segmentation or structuring of multimedia content and the recognition of low-level features within such content, comparatively little progress has been made on machine-generated semantic descriptions of audiovisual information.

The MPEG-7 standard [3] provides important functionalities for manipulation and management of multimedia content and its associated metadata. The extraction of semantic description and annotation of the content with the corresponding metadata though, is out of the scope of the standard, thus motivating heavy research efforts in the direction of automatic annotation of multimedia content. In order to make MPEG-7 accessible, re-usable and interoperable with many domains, the semantics of the MPEG-7 metadata terms need to be expressed in an ontology using a machine-understandable language. To this end, several approaches in the literature [4–7] address the problem of building multimedia ontologies to enable the inclusion and exchange of multimedia content through a common understanding of the multimedia content description and semantic information.

Acknowledging the importance of coupling domain-specific and low-level description vocabularies, we adopted a modular ontology infrastructure for the representation of the knowledge, to be used in a generic analysis scheme to semantically interpret and annotate multimedia content. The infrastructure consists of (i) a domain specific ontology that provides the necessary conceptualizations for the specific domain, (ii) multimedia ontologies that model the multimedia layer data in terms of low level features and media structure descriptors, and (iii) a core ontology (DOLCE [8]) that bridges the previous ontologies in a single architecture. Additionally, we present a semantic annotation tool, *M-OntoMat Annotizer*, that extends a previous framework for textual semantic annotation [9] and is capable of eliciting and representing knowledge both about content domain and the visual characteristics of multimedia data itself.

The purpose for constructing such a multimedia-targeted knowledge, beyond its own novelty, is to employ it in a image/video analysis framework to improve its performance. In this paper we describe a knowledge-assisted analysis (KAA) platform that manages to *bring in the loop* such an a-priori knowledge. The interaction between the analysis algorithms and the knowledge is continuous and tightly integrated, instead of being just a pre- or post-processing step in the overall architecture. To achieve this we used a region adjacency graph for image representation, that can interact dynamically (i.e. save, update, create new information) with the analysis processes. An initial segmentation algorithm generates a number of connected regions and then MPEG-7 visual descriptors are extracted for each region. A matching process queries the knowledge base and assigns each region a list of possible concepts along with a degree of relevance. Those concepts are used (among other information) for the construction of an RDF that is the actual system's output: A semantic interpretation of the multimedia in the formal syntax of an RDF.

The structure of this paper is as follows: Section 2 provides the general ontology infrastructure design, focusing on the multimedia related ontologies and structures. Section 3 describes the architecture and operation of the knowledge-assisted analysis platform, that was implemented in order to exploit the developed infrastructure to create automated semantic multimedia annotation. Conclusions are drawn and future directions are discussed in section 4.

2 Knowledge-base Infrastructure

Based on the above, we propose a comprehensive Ontology Infrastructure, the components of which will be described in this section. The challenge is that the hybrid nature of multimedia data must be necessarily reflected in the ontology architecture that represents and links multimedia and content layers. Fig. 1 summarizes the developed knowledge infrastructure.

Our framework uses *RDFS (Resource Description Framework Schema)* as modelling language. This decision reflects the fact that a full usage of the increased expressiveness of *OWL (Web Ontology Language)* requires specialized and advanced inference engines that are still not in mature state, especially when dealing with large numbers of instances with slot fillers.

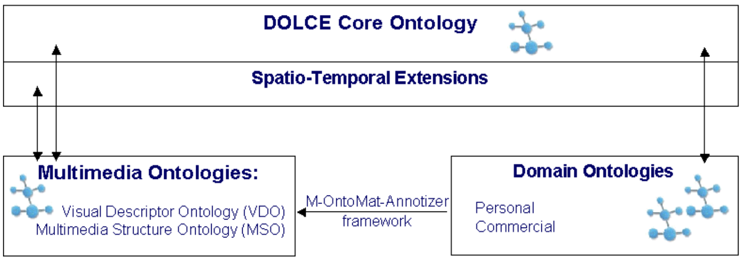

Fig. 1. Ontology Structure Overview

2.1 Ontology Infrastructure

The **Core Ontology**'s role 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* [8] 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.

Multimedia Ontologies model the domain of multimedia data, especially the visualizations in still images and videos in terms of low-level features and media structure descriptions. Structure and semantics are carefully modelled to be largely consistent with existing multimedia description standards like MPEG-7. Based on MPEG-7's Visual Part [10] and Multimedia Description Scheme [11], we have created the following two ontologies:

The *Visual Descriptor Ontology* (VDO) [12] contains the representations of the MPEG-7 visual descriptors and 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 (dominant color, region shape, etc). Although the construction of the VDO is tightly coupled with the specification of the MPEG-7 Visual Part, 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 *Multimedia Structure Ontology* (MSO) models basic multimedia entities from the MPEG-7 Multimedia Description Scheme and mutual relations like decomposition. MPEG-7 provides a number of tools for describing the structure of multimedia content in time and space. The Segment DS (Section 11 of [11]) describes a spatial and/or temporal fragment of multimedia content and a number of specialized subclasses are derived from that. These subclasses, that describe specific types of multimedia segments (such as video segments, moving regions, still regions and mosaics), along with their relations, have been modelled inside the MSO.

Domain Ontologies, in the multimedia annotation framework, 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. 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 multimedia becomes more efficient for the user 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.

2.2 M-OntoMat-Annotizer Framework

In order to exploit the ontology infrastructure presented above and annotate the domain ontologies with low-level multimedia descriptors the usage of a tool is necessary. The implemented framework is called *M-OntoMat-Annotizer* [1] (M stands for Multimedia) [13]. The development was based on an extension of the CREAM (CREAting Metadata for the Semantic Web) framework [9] and its reference implementation, *OntoMat-Annotizer*[2].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 components.

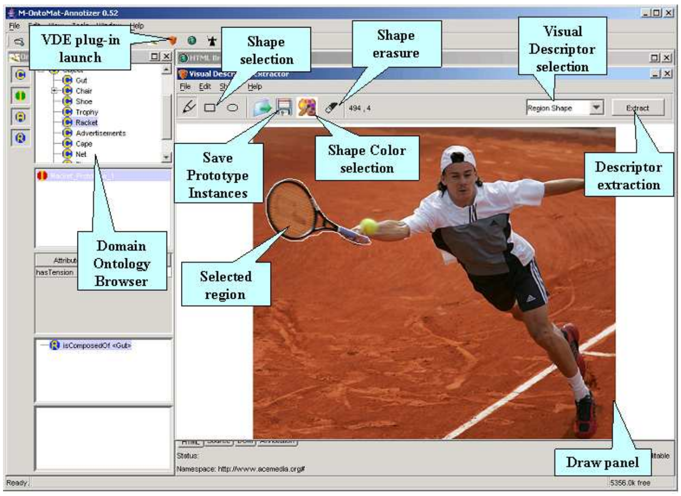

Fig. 2. The M-OntoMat-Annotizer user interface

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

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

The VDE plug-in facilitates the procedure of loading and processing visual content (images and videos), extracting visual feature and linking with domain ontology concepts. The interface, as shown in Fig. 2, seamlessly integrates with the common OntoMat-Annotizer interfaces. Usually, the user needs to extract the features (visual descriptors included in the VDO) 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-Annotizer ontology browser and selecting a region of interest the user can extract and link appropriate visual descriptor instances with domain concept prototype instances.

All the prototype instances are saved in an RDFS file. The VDE tool saves the domain concept prototype instances together with the corresponding descriptors, separately from the ontology file and leaves the original domain ontology unmodified. In this way, we manage to build the knowledge base that will serve as the primary reference resource for the multimedia content analysis process presented in the next section.

3 Knowledge-Assisted Multimedia Analysis

For the exploitation of the above described knowledge representation and the creation of content-based semantic metadata, we implemented a platform to test, measure and validate, firstly, the ontologies built and, secondly, the quality of the created semantic annotation. For this reason a prototype multimedia analysis system, named *Knowledge Assisted Analysis* (KAA) has been developed and its architecture and results are given in details in the sequent sections.

3.1 Platform Architecture

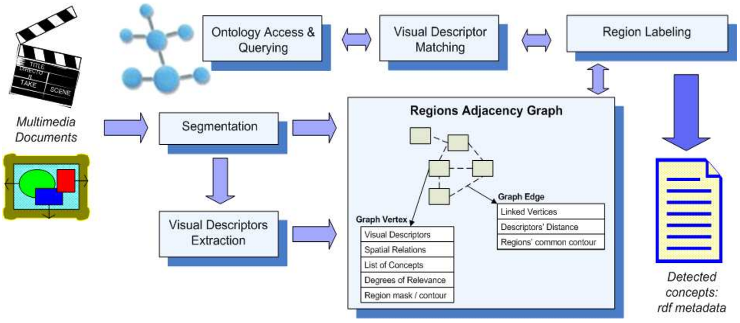

Fig. 3. Knowledge-Assisted Analysis architecture

KAA's general architecture scheme is depicted in Fig. 3, where in the center there is the Region Adjacency Graph, which is used as the representation of the image during analysis. Each graph's vertex corresponds to a connected region of the image whereas a graph's edge represents the link between two regions, holding the overall neighboring information. More specifically, in each vertex we store:

- the currently two supported MPEG-7 visual descriptors: Dominant Color and Region Shape
- the spatial relations between neighboring regions
- the list of possible concepts detected, along with a degree of relevance
- a list of all region's pixels (as it was found to be much more efficient than keeping a binary mask) and
- a list of all region's pixels belonging to the contour.

Likewise, in the graph's edge we keep information about:

- the linked regions
- the visual descriptors' distance of the linked regions and
- a list of pixels belonging in the common contour of the linked regions

After having defined a structure for image representation, we now introduce the algorithms that do the actual processing, interacting dynamically (i.e. initializing, updating, reading, etc) with the graph.

The first step should be a segmentation algorithm, that will actually provide a few tens of connected regions and initialize the graph. The segmentation used is an extension to the well known Recursive Shortest Spanning Tree (RSST) algorithm based on a new color model and so-called syntactic features [14]. The graph itself gives us the information whether two regions are neighboring or not, but we need to know additionally their spatial relation, e.g. region $\mathcal{X}$ is above of region $\mathcal{Y}$, or an absolute relation, like region $\mathcal{Z}$ is below all regions. Then, for each region we extract Dominant Color and Region Shape MPEG-7's visual descriptors and we store them in the corresponding vertex of the graph. The following step is to calculate for each region the distance between both its two descriptors and the corresponding descriptors of all prototype instances of all concepts in the VDO (explained in more details in section 3.2). The result of this matching is a distance for each descriptor, which is not very useful unless we find a way to produce a unique, combined distance. Towards this direction we tested both the simple approach of a weighted sum of the two distances and, to train a back-propagation neural network [15]. In this simple scenario of only two descriptors, both approaches worked fine. A typical normalization function is used and then the distance is inverted to degree of relevance, which is the similarity criterium for all matching and merging processes. From this whole procedure a list of possible concepts along with a degree of relevance for all regions is derived and stored appropriately in the graph.

In the case that two, or more neighboring regions have been assigned to only one concept, or other possible concepts have a degree less than a pre-defined threshold, we assume that these regions are part of a bigger region that wasn't segmented correctly due to the well-known segmentation limitations. We, then, correct this by merging all those regions, i.e. merging the graph's vertices and updating all the necessary graph's fields (extract again the visual descriptors, update region's contour, update graph's edges, etc).

3.2 Knowledge-base Retrieval

Whenever new multimedia content is provided as input for analysis, the existing a-priori knowledge base is used to compare, by means of matching the MPEG-7 visual descriptors, each region of the graph to the prototype instances of the multimedia domain ontologies. For this reason, the system needs to have full access to the overall knowledge base consisting of all domain concept prototype instances. These instances are applied as references to the analysis algorithms and with the help of appropriate rules related to the supported domains, KAA extracts semantic concepts that are linked to specific regions of the image or video shot.

For the actual retrieval of the prototypes and its descriptor instances, the *OntoBroker* [3] engine is used to deal with the necessary queries to the knowledge base. OntoBroker supports the loading of RDFS ontologies, so all appropriate ontology files can be easily loaded. For the analysis purposes, OntoBroker needs to load the domain ontologies where high-level concepts are defined, the VDO that contains the low-level visual descriptor definitions, and the prototype instances' files that include the knowledge base and provide the linking of domain concepts with descriptor instances. Appropriate queries are defined succeeding the retrieval of specific values from various descriptors and concepts. The OntoBroker's query language is *F-Logic* [4]. F-Logic is both a representation language that can model ontologies and a query language, so it can be used to query OntoBroker's knowledge.

3.3 Semantic Metadata Creation

The objective of this ontology-supported analysis, is to extract high level, human comprehensible features and create automatically semantic metadata describing the multimedia content itself. This metadata should also comply with a pre-defined format (so that in the future could be part of an annotation ontology) and for this reason the system's output is in RDF. For each image/video shot there is an RDF that contains a sequence of elements, one for each region/graph vertex. This element, as can be seen below in a small fragment of the resulting RDF, includes a list of candidate concepts with their degree of relevance and, additionally, information about the spatial relations with other regions.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:j.0="http://www.acemedia.org/ontologies/SCHEMA#"
  xmlns:j.1="http://www.acemedia.org/ontologies/BEACH-HOLIDAY#" >
  <rdf:Description rdf:about="http://www.acemedia.org/ontologies/INSTANCES#image1_segment3">
    <rdf:type rdf:resource="http://www.acemedia.org/ontologies/SCHEMA#StillRegion"/>
    <j.0:depicts rdf:resource="http://www.acemedia.org/ontologies/INSTANCES#Sea"/>
    <j.0:degree rdf:resource="http://www.acemedia.org/ontologies/INSTANCES#0.214404"/>
    <j.0:depicts rdf:resource="http://www.acemedia.org/ontologies/INSTANCES#Sand"/>
    <j.0:degree rdf:resource="http://www.acemedia.org/ontologies/INSTANCES#0.798639"/>
    <j.0:aboveOf rdf:resource="http://www.acemedia.org/ontologies/INSTANCES#image1_segment7"/>
    <j.0:rightOf rdf:resource="http://www.acemedia.org/ontologies/INSTANCES#image1_segment2"/>
    <j.0:aboveAll rdf:resource="http://www.acemedia.org/ontologies/INSTANCES#false"/>
    <j.0:belowAll rdf:resource="http://www.acemedia.org/ontologies/INSTANCES#false"/>
  </rdf:Description>
</rdf:RDF>
```

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

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

One could read this RDF and use it directly as semantic annotation by associating the specific image to the number of detected concepts. One step further would be to produce new concepts through the process of fuzzy reasoning (or any other form of reasoning) utilizing both the degrees and the spatial relations. Description Logics (DLs) have proved suitable for many applications (including multimedia) [16], since they have large expressive power while preserve low computational complexity. For that reason, DL based classical subsumption reasoning and rule-based reasoning with fuzzy extensions will be examined.

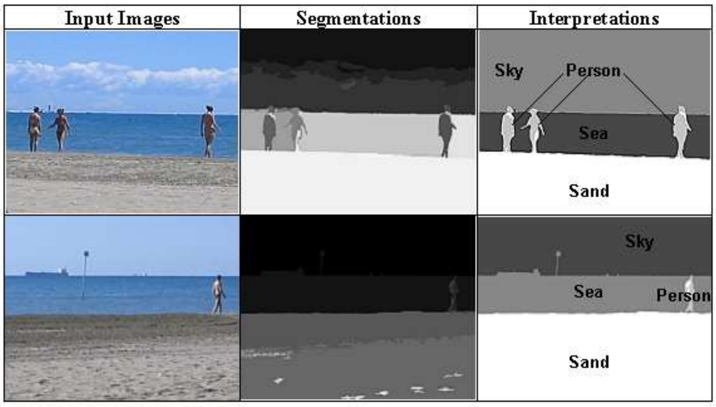

Fig. 4. Holiday-Beach domain results

As illustrated in Fig. 4, the resulting system's output includes also a segmentation mask outlining the semantic description of the scene. The different colors assigned to the generated regions correspond to concepts defined in the domain ontology. This labelled mask is nothing more than another representation of the concepts detected, without the strict format of an RDF, but with the major advantage of being very easily perceived by humans.

We conducted thorough experiments in the domain of beach and tennis having so far very promising results. In some cases when detection of specific concepts is difficult, then the user can evaluate himself the system's output by simply looking at the associated degree of relevance for each concept. In this way the user is provided a degree that measures the system's performance for a specific concept, or in other words, of how probable it is that this detected concept indeed describes correctly the image (or part of the image).

4 Conclusions and Future Work

In this paper we presented an integrated multimedia content annotation system that utilizes ontologies for the description of low-level visual features and for linking these descriptions to concepts in domain ontologies based on a prototype approach. The usually critical and time consuming procedure of prototype instances' construction is done with the help of a user-friendly tool, which hides analysis-specific details from the user. The major contribution of our work, in the area of semantic annotation of multimedia content, is the smooth integration of knowledge structures within image/video analysis processes. This has multiple benefits, such as visual descriptor evaluation, ontology refinement based on actual analysis results and finally, extension to other different kinds of images or video without changing the analysis algorithms, by simply employing different domain ontologies.

This work is still under evolution and many future extensions are under consideration. We focus our future work mainly towards three directions: (i) To improve the knowledge infrastructure, for example include partonomic relations of composite objects, add more prototype instances and process them in a more efficient way; (ii) integrate a reasoning engine that can refine the results in an iterative process and could also infer complicated concepts; (iii) use the graph structure even more dynamically by merging and splitting of regions according to a criterion of minimizing a distance function between regions and concepts available in the knowledge base. We believe that this approach is in the right direction for bridging the current semantic gap of content interpretation between humans and computers, which is the main hurdle for a wide expansion of multimedia in the Semantic Web.

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. 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.
2. J. Wielemaker A.Th. Schreiber, B. Dubbeldam and B.J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, May/June 2001.
3. 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.
4. 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.
5. R. Troncy. Integrating Structure and Semantics into Audio-Visual Documents. In *Proceedings of the 2nd International Semantic Web Conference (ISWC 2003)*, October 2003.
6. 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.
7. 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.

8. 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.
9. Siegfried Handschuh and Steffen Staab. Cream - creating metadata for the semantic web. *Computer Networks*, 42:579–598, AUG 2003. Elsevier.
10. ISO/IEC 15938-3 FCD Information Technology - Multimedia Content Description Interface - Part 3: Visual, March 2001, Singapore.
11. ISO/IEC 15938-5 FCD Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Scemes, March 2001, Singapore.
12. N. Simou, V. Tzouvaras, Y. Avrithis, G. Stamou, and S. Kollias. A Visual Descriptor Ontology for Multimedia Reasoning. In *In Proc. of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '05), Montreux, Switzerland, April 13-15, 2005.*, Montreux, Switzerland, April 13-15 2005.
13. 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.
14. 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.
15. E. Spyrou, H. Le Borgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O'Connor. Fusing MPEG-7 visual descriptors for image classiffication. In *Proc. of International Conference on Artificial Neural Networks ICANN 05*, Warsaw, Poland, September 2005.
16. U. Straccia. Reasoning within fuzzy description logics. *Journal of Artificial Intelligence*, 14:137–166, 2001.