# Semantic Image Analysis Optimization based on Context and Spatial Relations

**G. Th. Papadopoulos[1,2], Ph. Mylonas[3], V. Mezaris[2], Y. Avrithis[3] and I. Kompatsiaris[2]**

[1]Information Processing Laboratory, Electrical and Computer Engineering Department, Aristotle University of Thessaloniki
54124 Thessaloniki - Greece

[2]Informatics and Telematics Institute / Centre for Research and Technology Hellas, 1st Km Thermi-Panorama Road, 57001 Thessaloniki - Greece

[3]Image, Video and Multimedia Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens
15773 Zographou, Athens, Greece

## Abstract

In this chapter, we present our approach to semantic image analysis. Ontologies are used to capture a domain's general, spatial and contextual knowledge and a genetic algorithm is applied to fulfil the final annotation. The employed domain knowledge considers high-level information in terms of the concepts of interest of the examined domain, contextual information in the form of fuzzy ontological relations, as well as low-level information in terms of prototypical low-level visual descriptors. To account for the inherent ambiguities in visual information, uncertainty has been introduced and utilized within the spatial relations definition. To illustrate the proposed process, a hypotheses set of graded annotations is produced initially for each image region, and then context is exploited to update appropriately the estimated degrees of confidence. A genetic algorithm is applied as the last step, in order to select the most plausible annotation by utilizing the visual and spatial concept definitions that are included in the domain ontology. Experiments with a collection of photographs derived from two distinct domains demonstrate the performance of the proposed approach.

**Keywords:** semantic image analysis, knowledge-assisted analysis, multimedia ontologies, context, semantic annotation

# *1 Introduction*

Recent advances in both hardware and software technologies have resulted in an enormous increase in the number of images that are available in multimedia databases or over the Internet. As a consequence, the need for techniques and tools supporting their effective and efficient manipulation has emerged. To this end, several approaches have been proposed in the literature regarding the tasks of indexing, searching and retrieval of images. The very first attempts to address these issues concentrated on visual similarity assessment via the definition of appropriate quantitative image descriptions, which could be automatically extracted, and suitable metrics in the resulting feature space. Coming one step closer to treating images the way humans do, these were later adapted to a finer granularity level, making use of the output of segmentation techniques applied to the image (Smeulders, 2000). Whilst low-level descriptors, metrics and segmentation tools are fundamental building blocks of any image manipulation technique, they evidently fail to fully capture by themselves the semantics of the visual medium; achieving the latter is a prerequisite for reaching the desired level of efficiency in image manipulation. To this end, research efforts have concentrated on the semantic analysis of images, combining the aforementioned techniques with *a priori* domain specific knowledge, so as to result in a high-level representation of images (Al-Khatib, 1999). Domain specific knowledge is utilized for guiding low-level feature extraction, higher-level descriptor derivation, and symbolic inference.

One major obstacle, though, multimedia analysis still needs to overcome is the semantic gap (Mich, 1999; Smeulders, 2000); the latter forms an existing problem and in this approach we provide a partial contribution towards its solution. This hindrance becomes even harder when attempting to access vast amounts of multimedia information encoded, represented, and described in different formats and levels of detail. Although this gap has been acknowledged for a long time, multimedia analysis approaches are still divided into two main categories; the low-level multimedia analysis methods and tools on the one hand (e.g. (Milanese, 1993; Osberger, 1998; Oliva, 2001; Rapantzikos, 2005)) and the high-level semantic annotation methods and tools on the other hand (e.g. (Henderson, 1999; Tsechpenakis 2002; Benitez, 2003;

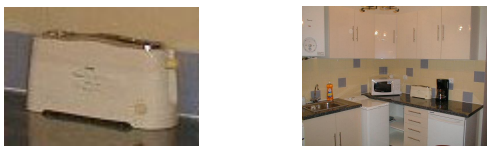
Voisine, 2005)). It was only recently, that state-of-the-art multimedia analysis systems have started using semantic knowledge technologies, as the latter are defined by notions such as the Semantic Web (Berners-Lee, 2001; W3C, Semantic Web, 2006) and ontologies (Gruber, 1993; Staab, 2004). The advantages of using Semantic Web technologies for the creation, manipulation and post-processing of multimedia metadata is depicted in numerous activities (Stamou, 2005), trying to provide "semantics to semantics".

Depending on the adopted knowledge acquisition and representation process, two types of approaches can be identified in the relevant bibliography: *implicit* ones, implemented by machine learning methods, and *explicit* ones, followed by model-based approaches. The use of machine learning techniques has proven to be a robust methodology for discovering complex relationships and interdependencies between numerical image data and the perceptually higher-level concepts, whereas they elegantly handle problems of high dimensionality, as well. Among the most commonly adopted machine learning techniques are Neural Networks (NNs), Hidden Markov Models (HMMs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Genetic Algorithms (GAs) (Mitchell, 1999; Zhang, 2001; Assfalg, 2005). On the other hand, model-based image analysis approaches make use of prior knowledge in the form of explicitly defined facts, models and rules, i.e., they provide a coherent semantic domain model to support "visual" inference in the specified context (Dasiopoulou, 2005; Hollink, 2005).

Regardless of the adopted approach towards knowledge representation, the inclusion of spatial information in the knowledge exploited during the analysis process demands the definition and extraction of spatial relations from the visual medium. The relevant literature considers roughly of two categories of approaches dealing with the latter task: *angle-based* and *projection-based* approaches. Angle-based approaches include (Wang, 2004), where a pair of fuzzy k-NN classifiers are trained to differentiate between the *Above-Below* and *Left-Right* relations, and the work of (Millet, 2005), where an individual fuzzy membership function is defined for every relation and applied directly to the estimated angle-histogram. Projection-based approaches include (Hollink, 2004), where qualitative directional relations in terms of

the centre and the sides of the corresponding objects' MBRs were defined, and (Skiadopoulos, 2005), where the use of a representative polygon was introduced.

Furthermore, it is rather true that in the real world, objects always exist in a context. In principle, a single image taken in an unconstrained environment is not sufficient to allow a computer algorithm or a human being to decide where an object starts and another object ends. However, a number of cues which are based on the statistics of our everyday's visual world are useful to guide this decision. This context may take the form of global image statistics which characterize an environment type, like an indoor office scene or an outdoor garden scene. Identification of an object in an image, or a close-up image of the same object may be difficult without being accompanied by useful contextual information.



(a) Isolated object (b) Object in context

Fig. 1. Isolated object vs. object in context.

As an example, an image of a horse is more likely to be present in a landscape environment such a green field, whereas a sofa is usually found indoors, or as depicted in Fig. 1, a close-up picture of a toaster is more difficult to identify or enroll when considered out of the rest environmental information of the image. Consequently, representing context is a research issue of great importance (Edmonds, 1999), affecting the quality of the produced results, especially in the field of multimedia analysis in general and knowledge-assisted image analysis in particular. The latter can be defined as a tightly coupled and constant interaction between low-level image analysis algorithms and higher-level knowledge representation (Athanasiadis, 2005); an area where the role of context is crucial. In recent years, a number of different context aspects related to image analysis have been studied, and a number of different approaches to model context representation have been proposed (Zhao, 1996; Mylonas, 2005; Mylonas, 2006).

Our work presents a radical at first sight approach to knowledge-assisted image analysis, based on coupling three independent components, such as explicit prior

knowledge (in the form of prototypical instances), spatial relations and contextual information. This approach is part of the aceMedia[1] EU-IST project dealing with efficient multimedia content access and personalized delivery. More specifically, a novel ontological representation for context is utilized, combining fuzzy theory and fuzzy algebra (Miyamoto, 1990; Klir, 1995) with characteristics derived from the Semantic Web, like the statement's reification technique (W3C, RDF Reification, 2004). In this process, confidence values of labels assigned to regions on the basis of low-level visual information similarity are optimized, according to a context-based confidence value readjustment algorithm (Mylonas, 2006). This is followed by a second optimization process, which utilizes the output of the former together with spatial information as input to a genetic algorithm, deciding on the optimal semantic interpretation of the image (Papadopoulos, 2006).

This chapter is organized in Sections as follows: Section 2 presents the overall aceMedia system architecture. Section 3 discusses low-level visual information processing, whereas Section 4 describes the employed knowledge infrastructure. Section 5 addresses the issues of context and spatial optimization, making use of the previously defined processing methods and knowledge representations. Experimental results for a collection of photographs belonging to two different domains are presented in Section 6 and some conclusions are drawn in Section 7.

# *2 System Overview*

## *2.1 Overall architecture*

The current approach was developed within the aceMedia project ([aceMedia]) and addresses the issues of efficient multimedia content access and personalized delivery by integration of multimedia analysis technologies with Semantic Web tools and techniques (Fig. 2). More specifically, aceMedia develops tools to automatically analyze content, generate semantic metadata and annotation, as well as support personalized and intelligent content search and retrieval services (Fig. 3).

[1] http://www.acemedia.org

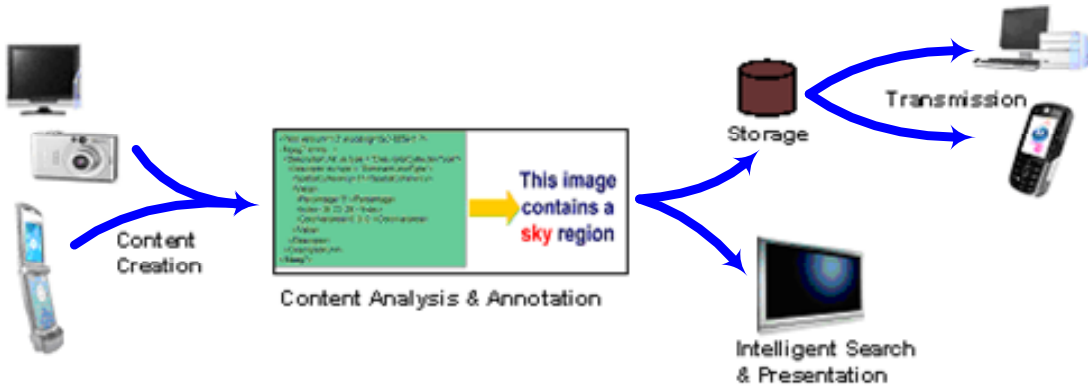



Fig. 2. Overview of the aceMedia system.

Key component of the aceMedia system is its Knowledge-Assisted Analysis module (KAA), which creates automatic multimedia annotations using an ontology driven approach. In KAA, low level image features are extracted from the multimedia content, using tools and techniques such as segmentation to atom regions and MPEG-7 descriptors extraction. Conversion of the MPEG-7 descriptors into an RDF (W3C, RDF, 2004) representation enables reasoning to be applied such that objects and areas in the scene can be identified, with reference to the appropriate domain ontology. Subsequently, the KAA module, using a methodology detailed in the sequel, decides on the labeling of the atom regions with a set of concepts from the domain ontology. The approach that is followed is generic and applicable to any domain, as long as appropriate domain ontologies are designed and made available.

Within aceMedia, the automatically generated metadata can be exploited by the personalization module which creates a model of user preferences and profiles enabling personalized search and presentation of content. The user model is dynamically updated by learning on user behavior as users interact with their content. Furthermore, semantic multimedia annotation is exploited in user centered applications, like intelligent search and retrieval. Latest aceMedia tools include user query interpretation, hybrid visual-semantic search and retrieval, and improved relevance feedback. In the remainder of this chapter, the focus will be on the Knowledge-Assisted Analysis module (KAA) of aceMedia and its supporting technologies.

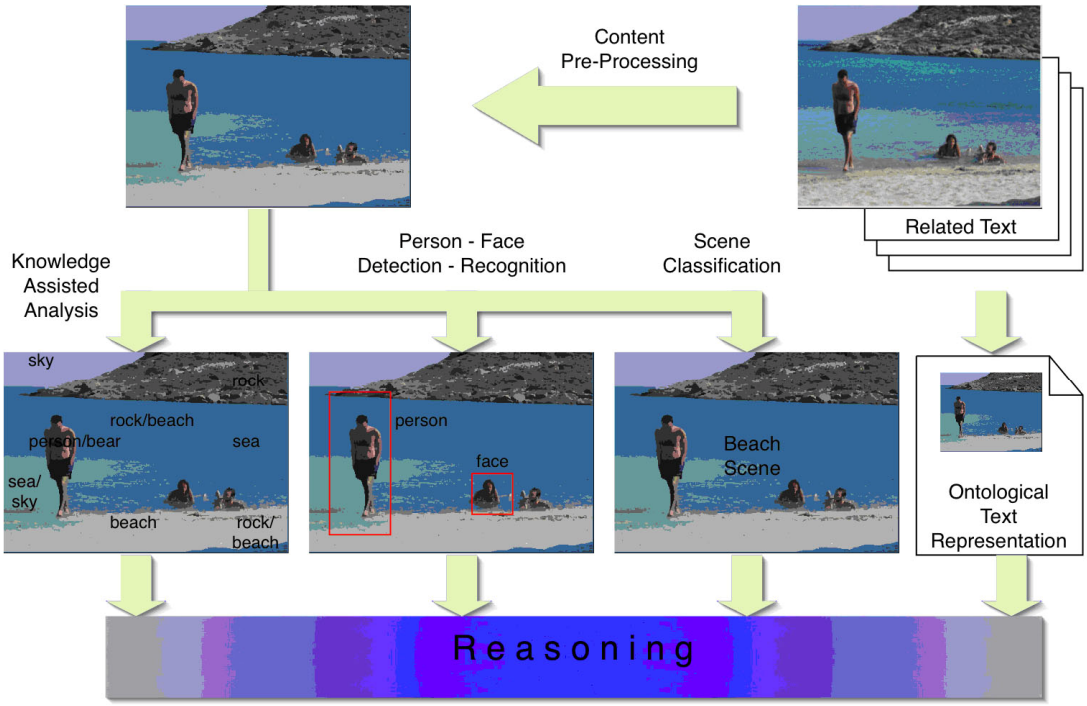



Fig. 3. Overall multimedia analysis and understanding architecture.

### *2.2 Knowledge assisted analysis within aceMedia*

The overall architecture of the proposed knowledge-assisted analysis framework is illustrated in Fig. 4. First segmentation is applied, and subsequently low-level descriptors and spatial relations are extracted for the generated image segments. Once the low-level descriptors are available, an initial set of hypotheses is generated for each image segment based on the distance between the segment's extracted descriptors and the domain concepts prototypical descriptors that are included in the knowledge base. Thereby, a set of plausible annotations (i.e., domain concepts) with corresponding degrees of confidence is produced for each segment. These graded hypotheses are then passed to the context analysis module that refines them utilizing the ad-hoc contextual knowledge, as will be described in more detail in the sequel. The refined hypotheses sets along with segment spatial relations are then passed to the genetic algorithm, which based on the provided domain concept definitions decides on the optimal semantic interpretation.

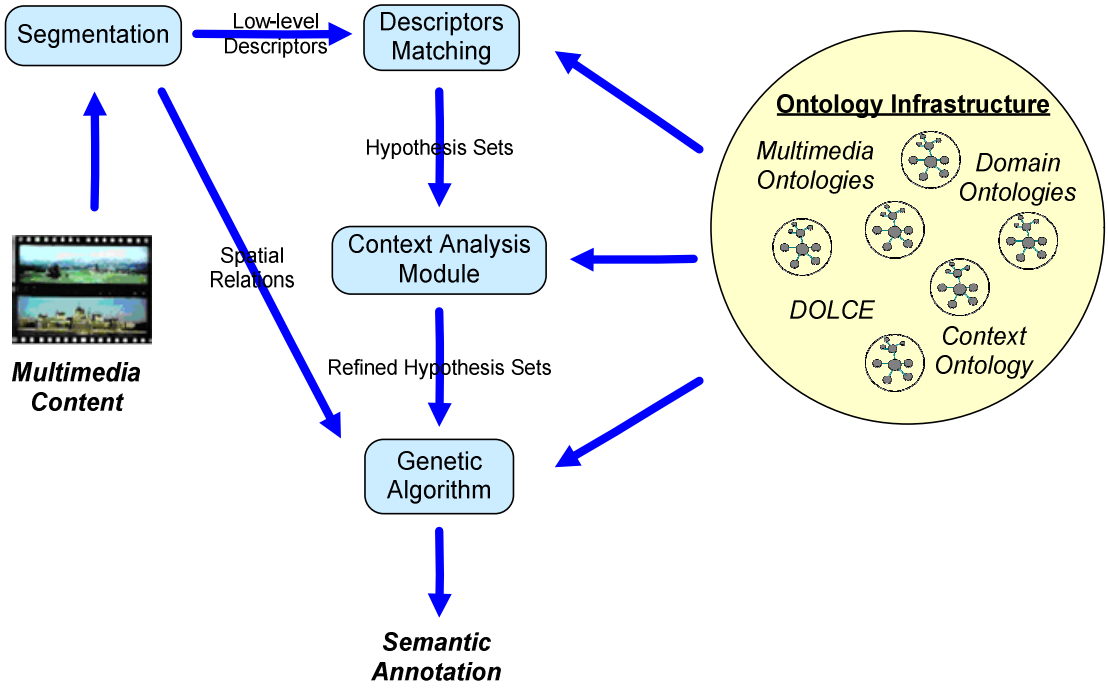



Fig. 4. Knowledge-assisted analysis framework architecture.

# 3 *Low-level Visual Information Processing*

## *3.1 Segmentation, feature extraction and initial hypotheses generation*

In order to implement the initial hypotheses generation procedure, the image under examination has to be segmented into regions and suitable low-level descriptions have to be extracted for every resulting segment. In the current implementation, an extension of the Recursive Shortest Spanning Tree (RSST) algorithm has been used for segmenting the image (Adamek, 2005). Considering the low-level descriptions, a specific set of descriptors defined within the MPEG-7 standard have been selected, namely the *Homogeneous Texture*, *Region Shape* and *Dominant Colour* descriptors. Their actual extraction, when dealing with each and everyone of the generated image regions, is performed according to the guidelines provided by the MPEG-7 eXperimentation Model (XM) (MPEG-7 Visual Experimentation Model (XM), 2001).

In order to produce the hypotheses sets, appropriate measures need to be defined for qualitatively assessing visual similarity between the examined image segments and the defined domain concept prototypes. As MPEG-7 does not provide a standardized

method for combining different descriptors distances or for estimating a single distance based on more than one descriptor, a weighted sum approach was followed, resulting in the calculation of a single scalar distance $D$ for each hypothesis. Thereby, a similarity degree $DOC$ is produced per segment against each of the defined domain concepts, as follows:

$$DOC = \frac{1}{e^{mD}} \qquad (1)$$

where the slope parameter $m$ is experimentally set. The pairs of each domain concept and its corresponding degree of confidence that result for each image segment comprise its initial hypotheses set.

### *3.2 Fuzzy spatial relations extraction*

Exploiting domain-specific spatial knowledge in image analysis tasks is a common practice among the object recognition community. It is generally observed that objects tend to be present in a scene within a particular spatial context and thus spatial information can substantially assist in discriminating between objects exhibiting similar visual characteristics. The use of spatial context forms the key for the unambiguous recognition process, as it refers to the relationships among the location of different objects in the scene; spatial context is associated to spatial relationships between objects or regions in a still image or video sequence. In general, at least two types of meaningful spatial contextual relationships can be identified in natural images. First, relationships exist between spatial co-occurrence of certain objects in natural images. For example, repeated detection of *snow* would imply low *grass* probability. Second, relationships exist between spatial locations of certain objects within an image: *grass* tends to occur below *sky*, *sky* above *snow*, etc. Of course, the set of spatial relationships can be rich (many spatial relationships with minor differences between each) or sparse (fewer distinct relationships). The spatial relations define the absolute or relative spatial information between objects. Among the most commonly adopted spatial relations, *directional* ones have received particular attention. In the present analysis framework, eight fuzzy directional relations are supported, namely *Above* (*A*), *Right* (*R*), *Below* (*B*), *Left* (*L*), *Below-Right* (*BR*), *Below-Left* (*BL*), *Above-Right* (*AR*) and *Above-Left* (*AL*).

In the proposed analysis approach, the extraction of fuzzy directional relations builds on the principles of projection- and angle- based methodologies (Skiadopoulos, 2005; Wang, 2004) and can be decomposed in the following steps. First, a *reduced box* is computed from the *ground* object's (i.e., the object used as reference, painted dark grey in Fig. 5) Minimum Bounding Rectangle (MBR), so as to include the object in a more representative way. The computation of this *reduced box* is performed in terms of the MBR compactness value *c*, which is defined as the value of the fraction of the object's area to the area of the respective MBR: if the initially computed *c* is below a threshold *T*, the ground object's MBR is reduced repeatedly until the desired threshold is satisfied. Then, eight cone-shaped regions are formed on top of this reduced box, as illustrated in Fig. 5, each corresponding to one of the defined directional relations. The percentage of the *figure* object's (i.e., the object whose relative position is to be estimated, painted light grey in Fig. 5) pixels that are included in each of the cone-shaped regions determines the degree to which the corresponding directional relation is satisfied. After extensive experimentations, the value of threshold *T* was set equal to 0.85.

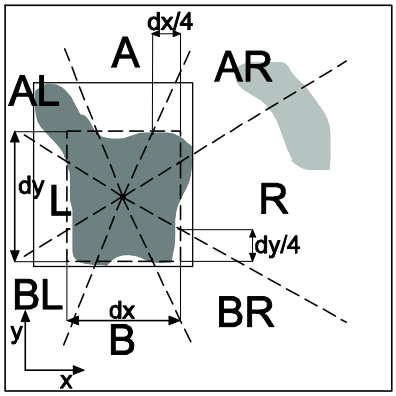



Fig. 5. Fuzzy spatial relation definition – Minimum Bounding Rectangle (MBR).

# *4 Knowledge Infrastructure*

Among the possible knowledge representation formalisms, ontologies present a number of advantages (Gruber, 1993; Staab, 2004). They provide a formal framework for supporting explicit, machine-processable semantics definitions, and they facilitate inference and the derivation of new knowledge based on a set of rules, as well as already existing knowledge. Thus, ontologies are suitable for expressing multimedia content semantics in a formal machine-processable representation that will allow

automatic analysis and further processing of the extracted semantic descriptions. Following these considerations, in the aceMedia project framework has been used an RDF-based ontology infrastructure, introduced in (Bloehdorn, 2005), as the means for representing the necessary knowledge components. As illustrated in Fig. 6, this knowledge representation consists of (i) a *Core Ontology*, whose role is to serve as a starting point for the construction of new ontologies, (ii) a *Visual Descriptor Ontology*, that contains the representations of the MPEG-7 visual descriptors, (iii) a *Multimedia Structure Ontology*, that models basic multimedia entities from the MPEG-7 Multimedia Description Scheme (ISO/IEC Part:3, 2001), and (iv) a set of *Domain Ontologies*, that model the content layer of multimedia content with respect to specific real-world domains. In the following of this Section, we shall briefly examine each one of them, focusing on the necessary details where needed.

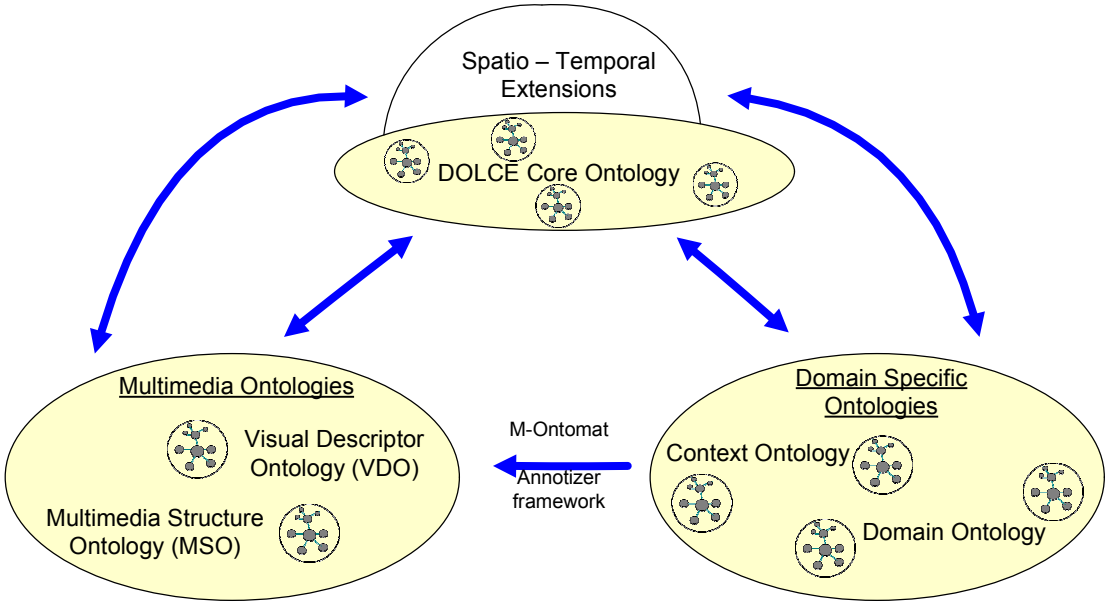



Fig. 6. RDF-based knowledge infrastructure.

### *4.1 Core ontology*

In general, core ontologies are typically conceptualizations that contain specifications of domain-independent concepts and relations based on formal principles derived from philosophy, mathematics, linguistics, and psychology. The role of the core ontology in this overall framework is to serve as a reference 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 other existing ontologies within the architecture. In the presented framework, the *DOLCE* (Gangemi, 2002)

ontology is used for this purpose. DOLCE was explicitly designed as a core ontology, is minimal in the sense that it includes only the most reusable and widely applicable upper-level categories, rigorous in terms of axiomatization, and extensively researched and documented. On top of DOLCE and in order to further accommodate the corresponding directional and topological relationships in the spatial domain, concepts taken from the 'Region Connecting Calculus' (Cohn, 1997), Allen's interval calculus (Allen, 1983) and directional models (Papadias, 1997; Skiadopoulos, 2004) have been carefully incorporated (Simou, 2005a), included in Fig. 6 as the set of DOLCE's spatio-temporal extensions.

### *4.2 Visual Descriptor Ontology*

The visual descriptor ontology (VDO) (Simou, 2005) represents the visual part of the MPEG-7 and thus, contains the representations of the set of visual descriptors used for knowledge-assisted analysis. Its modelled concepts and properties describe the visual characteristics of the objects. The construction of the VDO attempted to follow the specifications of the MPEG-7 Visual Part (ISO/IEC Part:3, 2001), however, because a strict attachment to the MPEG-7 Visual Part became impossible, several requisite modifications were made in order to adapt the XML schema provided by MPEG-7 to an ontology and the data-type representations available in RDFS. The tree of the VDO consists of four main concepts, namely: *VDO:Region*, *VDO:Feature*, *VDO:VisualDescriptor* and *VDO:Metaconcepts*. It should be noted that none of these concepts is included in the XML schema defined in MPEG-7, but their need was crucial in order to create a correctly defined ontology. The *VDO:VisualDescriptor* concept contains the visual descriptors, as these are defined by MPEG-7. The *VDO:Metaconcepts* concept on the other hand, contains some additional concepts that were necessary for the VDO, but they are not clearly defined in the XML schema of MPEG-7. The definition of the remaining two concepts *VDO:Region* and *VDO:Feature* was necessary in order to enable the linking of visual descriptors to the actual image regions. For instance, let us consider the *VDO:VisualDescriptor* concept, which consists of six subconcepts, one for each category of the MPEG-7-specified visual descriptors. These are: *colour*, *texture*, *shape*, *motion*, *localization*, and *basic descriptors*. Each of these subconcepts includes a number of relevant descriptors that are defined as concepts within the VDO.

### *4.3 Multimedia structure ontology*

The multimedia structure ontology (MSO) models basic multimedia entities from the MPEG-7 Multimedia Description Scheme (ISO/IEC Part:5, 2001) and mutual relations like *decomposition*. Multimedia content is considered to be classified into five types within the MPEG-7 standard, each of which has its own segment subclasses, namely: *image*, *video*, *audio*, *audiovisual*, and *multimedia*. The standard provides a variety of tools for describing the structure of multimedia content. A spatial or temporal fragment of multimedia content is described by the Segment DS (ISO/IEC Part:5, 2001). More specifically, a number of specialized subclasses are derived from it, that describe the specific types of multimedia segments (such as: video segments, moving regions, still regions, mosaics, etc.) resulting from spatial, temporal, and spatiotemporal segmentation of the different multimedia content types. It should be also stressed out, that multimedia resources may be segmented into sub-segments through four types of decomposition, namely: *spatial*, *temporal*, *spatiotemporal*, and *media source*.

### *4.4 Domain ontology*

A domain ontology was developed for representing the knowledge components that need to be explicitly defined under the proposed approach. This contains the semantic concepts that are of interest in the examined domain (e.g., in the beach vacation domain: Sea, Sand, Person, etc.), their prototypical low-level characteristics, as well as their spatial relations.

As opposed to concepts themselves that are manually defined by domain experts, prototypical visual descriptor instances for each of the concepts of interest, which are required for the initial hypotheses generation during the matching process described in section 3.1, and spatial relations, are extracted using a training set of images. More specifically, to populate the domain knowledge with prototypical visual descriptor instances, sample images of a training set are processed with the M-Ontomat-Annotizer tool (Bloehdorn, 2005), that allows linking domain concepts with low-level visual descriptor values (Saathoff, 2006). The values of spatial relations for the concepts of the given domain are estimated according to the following ontology population procedure:

Let $S = \{s_i, i = 1,\ldots I\}$ denote the set of regions produced for the image under consideration by the segmentation process, $C = \{c_p, p = 1,\ldots P\}$ denote the set of concepts defined in the employed domain ontology and

$$\Pi = \{\rho_k, k = 1,\ldots K\} = \{A,\ AL,\ AR,\ B,\ BL,\ BR,\ L,\ R\} \qquad (2)$$

denote the set of supported spatial relations. Then, the degree to which $s_i$ satisfies relation $\rho_k$ with respect to $s_j$ can be denoted as $I_{\rho_k}(s_i, s_j)$, where the values of function $I_{\rho_k}$ are estimated according to the fuzzy spatial relations extraction procedure of Section 3.2 and thus belong to the [0, 1] interval. To populate the ontology, this function needs to be evaluated over a set of segmented images with ground truth annotations that serves as a training set. More specifically, the mean values, $I_{\rho_k mean}$, of $I_{\rho_k}$ are estimated, for every $k$ over all region pairs of segments assigned to objects $(c_p, c_q)$, $p \neq q$. The calculated values are stored in the ontology. These constitute the constraints input to the spatial optimization problem which is solved by the genetic algorithm, as will be described in Section 5.2.

### *4.5 Context Ontology*

#### *4.5.1 A "fuzzified" context model*

As found in the literature, the term *context* has been widely studied and has many interpretations, as well as definitions (Mylonas, 2005), none of which is globally applicable. It is therefore very important to establish a working interpretation for context, in order to benefit from and contribute to multimedia analysis. The ultimate goal is to develop a non-scene specific method for generating context models useful for general scene understanding. The problems to be addressed in this Section include how to represent context, how to determine it, and how to use it to optimize the results of knowledge-assisted analysis. Results of the latter are highly dependent on the domain an image belongs to and thus in many cases are not sufficient for the understanding of multimedia content. The lack of contextual information (Mylonas, 2005) in the above process is a major limitation towards a better analysis performance and together with similarities in numerous low-level characteristics of various object types (such as: *colour*, *texture*, *shape*, etc.) results in a significant number of

misclassifications. Herein, we introduce a method for further improving the results of the proposed knowledge-based approach, based on a contextual ontology and focus on its knowledge representation, as its role is crucial in the understanding of the context optimization process that follows in Section 5.1.

In general, it is possible to formally describe an ontology as the entire set of concepts and semantic relations between concepts within a given universe:

$$O = \{C, \{R_{c_i,c_j}\}\},\ \ i,j = 1,\ldots n\ ,\ R_{c_i,c_j} : C \times C \to \{0,1\},\ \ i,j = 1,...n \qquad (3)$$

where $O$ forms an ontology, $C$ is the set of all possible concepts it describes and $R_{c_i,c_j}$ denotes the semantic relation amongst two concepts $c_i, c_j$. Any type of relation may be included in an ontology, however, for the problem at hand a "fuzzified", ad-hoc context ontology is introduced. In order for this ontology to be highly descriptive and accurate, it must contain a representative number of distinct and even diverse relations among concepts, so as to scatter information among them and meaningfully describe context. In this work we utilize a set of relations, whose semantics are defined in MPEG-7 (Benitez, 2001; ISO/IEC Part:5, M 4242, 2001), namely: partOf (*P*), specializationOf (*Sp*), propertyOf (*Pr*), inContextOf (*Ct*), locationOf (*Loc*), instrumentOf (*Ins*) and patientOf (*Pat*).

However, when modelling real-life information governed by uncertainty and fuzziness, only fuzzy relations can handle such issues. In fact, the above commonly encountered relations can be modelled as fuzzy relations. Thus, in order to extract and use the desired ontological context, we define it by means of fuzzy ontological relations:

$$O_F = \{C, \{r_{c_i,c_j}\}\},\ \ i,j = 1,...n \qquad (4)$$

where $O_F$ forms a domain-specific "fuzzified" ontology, $C$ is the set of all possible concepts it describes, $r_{c_i,c_j} = F(R_{c_i,c_j}) : C \times C \to [0,1],\ R_{c_i,c_j} : C \times C \to \{0,1\}\ ,i,j = 1,...n$ denotes a fuzzy ontological relation amongst two concepts $c_i, c_j$ and $R_{c_i,c_j}$ is a crisp semantic relation amongst the two concepts. We shall use this "fuzzified" definition of the knowledge model throughout the rest of this chapter.

### *4.5.2 Contextual knowledge representation and ontological relations*

The proposed contextual ontology model is able to represent any type of fuzzy relation between concepts $F(R_{c_i,c_j}) = r_{c_i,c_j}$. All relations between concepts are contained within an RDF-based representation, forming the overall contextual knowledge. Describing the accompanying degree of confidence is carried out using reification (W3C, RDF Reification, 2004), i.e., by making a statement about the statement, which contains the degree information. Reification was used in order to achieve the desired expressiveness, whereas representing fuzziness with reified statements is an acceptable way, since the reified statement should not be asserted automatically. For instance, having a statement, such as *Car inContextOf MotorsportScene* and a degree of confidence of *0.85* for this statement, does obviously not entail, that a car is always in the context of a motorsports scene.

To illustrate things further, let us select one fuzzy relation, e.g., the *partOf* relation *P*, which, according to the previous analysis, is a fuzzy taxonomic relation on the set of concepts. $P(a,b) > 0$ means that *b* is a part of *a*. For instance, *a* could be a *boat* and *b* could be a *sail*. An example of its formal representation is presented in Fig. 7.

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:context="&dom;"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
    <rdf:Description rdf:about="#partOf">
        <rdfs:domain>
            <rdf:Description rdf:about="http://www.w3.org/1999/02/22-rdf-syntax-ns#Statement"/>
        </rdfs:domain>
        <rdfs:range>
            <rdf:Description rdf:about="http://www.w3.org/2001/XMLSchema#float"/>
        </rdfs:range>
    </rdf:Description>
    <rdf:Description rdf:about="#relation1">
        <rdf:subject rdf:resource="&dom;sail"/>
        <rdf:predicate rdf:resource="&dom;partOf"/>
        <rdf:object> rdf:resource="&dom;boat"</rdf:object>
        <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Statement"/>
        <context:partOf
rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.85</context:partOf>
    </rdf:Description>
</rdf:RDF>
```

Fig. 7. Reified RDF/XML representation of the *partOf* fuzzy relation.

The proposed model can be visualised as a graph, in which every node represents a concept and each edge between two nodes a contextual relation between the respective concepts. Additionally each edge has a corresponding degree of confidence that represents the fuzziness that exists within the context model. Non-existing edges are implying non-existing relations, i.e. relations with zero confidence values are omitted.

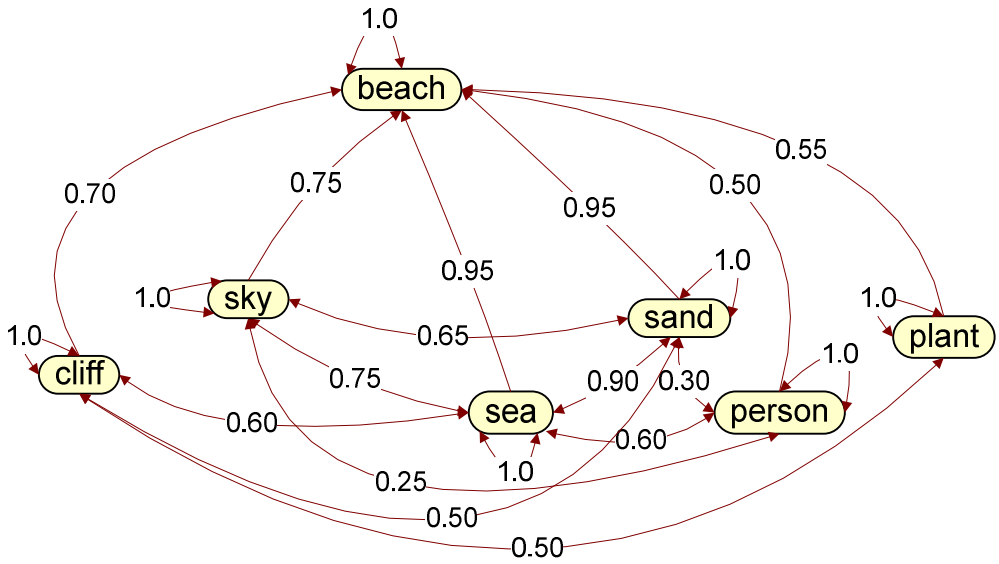



Fig. 8. *locationOf* ontology fragment from the beach vacation domain.

Finally, another important point to consider is the fact that each concept has a different probability to appear in the scene. A flat context model, i.e., relating concepts only to the respective scene type, would not be sufficient in this case. We model a more detailed graph, where ideally concepts are all related to each other, implying that the graph relations used are in fact transitive. As observed in Fig. 8, every concept participating in the contextualized ontology has at least one link to the *root element*. Additional degrees of confidence exist between any possible connections of nodes in the graph, whereas the *root element beach* could be related either directly or indirectly with any other concept. This results to the notion of *context relevance*, described in greater detail in the following section of this work.

# 5 Context and Spatial Optimization

## 5.1 Context optimization

Once the contextual knowledge structure is defined and the corresponding representation is implemented, a context-based confidence value readjustment algorithm is introduced to aid the scope of multimedia analysis. Our contextualization approach acts as a post-processing step on top of the initial hypotheses set and re-estimates the initial degree of confidence of each label for each image segment. In the process, it utilizes contextual information residing in the aforementioned context ontology and passes the optimized results as input to the genetic algorithm. We exploit a contextual form constructed by a semantically meaningful combination of the previously selected fuzzy relations. More specifically, each segment's *label* is related to a specific *concept* $c_k$ of the application domain ontology and stored together with its relationship degrees to any other related concept. To tackle cases that more than one concept is related to multiple concepts, we introduce the term *context relevance* $cr_{dm}(c_k)$, which refers to the overall relevance of concept $c_k$ to the *root element* of the domain $dm$. An exhaustive approach, that considers all possible routes in the graph, is followed, with respect to the fact that all routes between concepts are reciprocal at large.

Estimation of each concept's context relevance is derived from two sources, namely from:

1. *direct relationships* of the concept with other concepts and
2. *indirect relationships*, utilizing a suitable distance metric operator.

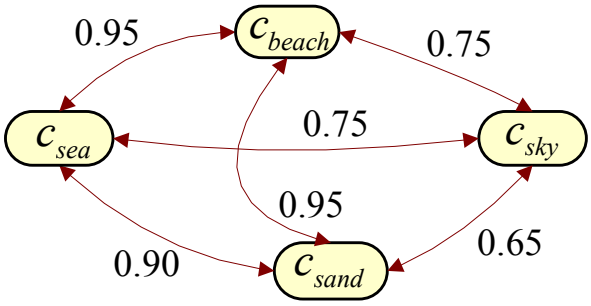



Fig. 9. Simplified context ontology graph sample.

To the aid of the above, let us present (Fig. 9) a simplified, hence illustrative example of a concept's context relevance calculation, derived from the beach vacation contextualized ontology part, presented in Fig. 8, assuming that the only available concepts were $c_{beach}$, $c_{sea}$, $c_{sand}$ and $c_{sky}$. Let concept $c_{sea}$ be related to concepts $c_{beach}$, $c_{sky}$ and $c_{sand}$ directly with: $r_{c_{sea},c_{beach}} = 0.95$, $r_{c_{sea},c_{sky}} = 0.75$ and $r_{c_{sea},c_{sand}} = 0.90$, while concept $c_{sky}$ is related to concept $c_{beach}$ with $r_{c_{sky},c_{beach}} = 0.75$ and to concept $c_{sand}$ with $r_{c_{sky},c_{sand}} = 0.65$ and concept $c_{sand}$ is additionally directly related to concept $c_{beach}$ with $r_{c_{sand},c_{beach}} = 0.95$. Given the semantic perspective on the correlation between any two concepts, we select the *max* operator as the appropriate distance metric operator. Then, we calculate the value for $cr_{beach}(c_{sky})$ as follows:

$$
\begin{aligned}
& cr_{beach}(c_{sky}) \\
& = \max\left\{ r_{c_{sky},c_{beach}}, r_{c_{sky},c_{sea}} \times r_{c_{sea},c_{beach}}, r_{c_{sky},c_{sand}} \times r_{c_{sand},c_{beach}}, r_{c_{sky},c_{sand}} \times r_{c_{sand},c_{sea}} \times r_{c_{sea},c_{beach}}, r_{c_{sky},c_{sea}} \times r_{c_{sea},c_{sand}} \times r_{c_{sand},c_{beach}} \right\} \\
& = \max\{0.75, 0.7125, 0.6175, 0.55575, 0.64125\} \\
& = 0.75
\end{aligned}
$$

In this case, we observe that the direct relationship between the two concepts dominates the context relevance value for concept *sky*. The reader is encouraged to assume that a similar approach is followed for every concept participating in the context ontology.

After estimating each concept's context relevance value and according to the contextualization algorithm described in (Mylonas, 2006), we identify the optimal normalization parameter for the domain at hand and define a minimum considerable value for any potential degree of confidence. The meaning of this normalization parameter lies in the fact that our algorithm only examines labels accompanied by a degree of confidence higher than this value, i.e., we examine the supplied domain ontology and identify the concept in the domain that is related to it, only if the initial degree lies above this experimentally identified threshold, in order to reduce the redundancy and computational complexity. Then for each identified concept we obtain the particular contextual information in the form of its relations to the set of any other concepts and calculate the new degree of confidence for the label associated to the region, based on the normalization parameter and the context's relevance value. In the case a concept is related to additional concepts apart from the *root element* of

the ontology, an intermediate aggregation step is applied to calculate the concept's context relevance value, as already explained.

Key points in this approach are the identification of the inter-concept relationships between all concepts, the definition of a meaningful normalization parameter and the identification of the optimal initialization value for the initial confidence values. When (re-)evaluating these values, the ideal normalization parameter is always defined with respect to the particular domain of knowledge and is the one that quantifies the semantic correlation to the domain. The overall process is terminated when belief to the labelling output provided initially is not strong enough, i.e., there are no more labels with an acceptable initial confidence value above the specified initialization value. The result of this contextualization step is the meaningful readjustment of the initial degrees of confidence accompanying each image segment, increasing the efficiency and robustness of the proposed semantic image analysis methodology and providing optimized input to the genetic algorithm, as described in the next Section.

### *5.2 Spatial Optimization*

As outlined in Section 2 of this chapter, after the initial set of hypotheses is generated (based solely on visual features) and refined using context, a genetic algorithm (GA) is introduced to decide on the optimal image interpretation using the fuzzy spatial relations that have been computed for every pair of image segments. The GA is employed to solve a global optimization problem, while exploiting the available domain spatial knowledge, and thus overcoming the inherent visual information ambiguity. Spatial knowledge is obtained according to the guidelines of Section 4.4 and the resulting learnt fuzzy spatial relations serve as constraints denoting the allowed domain objects spatial topology.

#### ***5.2.1*** *Fitness function*

The proposed optimization process utilizes as input (i) the context-refined hypotheses sets (as already described in Section 5.1), (ii) the fuzzy spatial relations extracted between the examined image segments, and (iii) the spatial-related domain knowledge as produced by the particular training process. Under the proposed approach, each

chromosome represents a possible solution. Consequently, the number of the genes comprising each chromosome equals the number *I* of the segments $s_i$ produced by the segmentation algorithm and each gene assigns a defined domain concept to an image segment.

An appropriate *fitness function* is introduced to provide a quantitative measure of each solution's fitness, i.e. to determine the degree to which each interpretation is plausible:

$$f(CR) = \lambda \cdot FS_{norm} + (1-\lambda) \cdot SC_{norm} \quad (5)$$

where *CR* denotes a particular chromosome, $FS_{norm}$ refers to the degree of low-level descriptors matching, and $SC_{norm}$ stands for the degree of consistency with respect to the provided spatial domain knowledge. The variable $\lambda$ is introduced to adjust the degree to which visual features matching and spatial relations consistency should affect the final outcome.

The value of $FS_{norm}$ is computed as follows:

$$FS_{norm} = \frac{\sum_{i=1}^{N} I_M(g_{ip}) - I_{\min}}{I_{\max} - I_{\min}} \quad (6)$$

where:

$$I_M(g_{ip}) \equiv DOC_{ip} \quad (7)$$

denotes the degree to which the visual descriptors extracted for segment $s_i$ match the ones of concept $c_p$, and where $g_{ip}$ represents the particular assignment of $c_p$ to $s_i$. Thus, $I_M(g_{ip})$ gives the degree of confidence, $DOC_{ip}$ (as defined in Section 3.1), associated with each hypothesis. $I_{\min} = \sum_{i=1}^{N} \min_p I_M(g_{ip})$ is the sum of the minimum degrees of confidence assigned to each region hypotheses set and $I_{\max} = \sum_{i=1}^{N} \max_p I_M(g_{ip})$ is the sum of the maximum degrees of confidence values respectively. For the computation of $SC_{norm}$ the approach described in the following subsection 5.2.2 is followed.

### *5.2.2* *Spatial constraints verification*

The exploitation of spatial information in the analysis procedure relies on the estimation of the degree to which the spatial constraints between two objects are satisfied for a pair of object- segment mappings $g_{ip}$, $g_{jq}$. In this work, this degree of satisfaction is expressed by the function $I_S(g_{ip}, g_{jq})$, which is defined with the help of a normalized Euclidean distance $d(g_{ip}, g_{jq})$. The latter is calculated according to the following equation:

$$d(g_{ip}, g_{jq}) = \frac{\sqrt{\sum_{k=1}^{8}\left(I_{\rho_k mean}\left(c_p, c_q\right) - I_{\rho_k}\left(s_i, s_j\right)\right)^2}}{\sqrt{8}} \qquad (8)$$

where $I_{\rho_k mean}$ is part of the knowledge infrastructure, as discussed in Section 4, $I_{\rho_k}(s_i, s_j)$ denotes the degree to which spatial relation $\rho_k$ is verified for a certain pair of segments $s_i, s_j$ of the examined image and $c_p, c_q$ denote the domain defined concepts assigned to them respectively. Distance $d(g_{ip}, g_{jq})$ receives values in the interval [0, 1]. Consequently, the function $I_S(g_{ip}, g_{jq})$ is then defined as:

$$I_S(g_{ip}, g_{jq}) = 1 - d(g_{ip}, g_{jq}) \qquad (9)$$

and takes values in the interval [0, 1] as well, where 1 denotes an allowable relation and 0 denotes an unacceptable one. Using this, the value of $SC_{norm}$ is computed according to the equation:

$$SC_{norm} = \frac{\sum_{l=1}^{W} I_{S_l}\left(g_{ij}, g_{pq}\right)}{W} \qquad (10)$$

where *W* denotes the number of the constraints that had to be examined.

### *5.2.3* *Implementation issues*

To implement the previously described optimization process, a population of 200 chromosomes is employed, and it is initialized with respect to the input set of

hypotheses. After the population initialization, new generations are iteratively produced until the optimal solution is reached. Each generation results from the current one through the application of the following three operators.

- *selection*: a pair of chromosomes from the current generation are selected to serve as parents for the next generation. In the proposed framework, the Tournament Selection Operator (Goldberg, 1991), with replacement, is used.
- *crossover*: two selected chromosomes serve as parents for the computation of two new offsprings. Uniform crossover with probability of 0.7 is used.
- *mutation*: every gene of the processed offspring chromosome is likely to be mutated with probability of 0.008. If mutation occurs for a particular gene, then its corresponding value is modified, while keeping unchanged the degree of confidence.

Parameter $\lambda$, regulating the relative weights of low-level descriptor matching and spatial context consistency was set to 0.35 after experimentation. The resulting weight of $SC_{norm}$, points out the importance of spatial context in the optimization process.

To ensure that chromosomes with high fitness will contribute to the next generation, the overlapping populations approach was adopted. More specifically, assuming a population of $m$ chromosomes, $m_s$ chromosomes are selected according to the employed *selection* method, and by application of the *crossover* and *mutation* operators, $m_s$ new chromosomes are produced. Upon the resulting $m + m_s$ chromosomes, the *selection* operator is applied once again in order to select the $m$ chromosomes that will comprise the new generation. After experimentation, it was shown that choosing $m_s = 0.4 \cdot m$ resulted in higher performance and faster convergence. The above iterative procedure continues until the diversity of the current generation is equal to/less than 0.001 or the number of generations exceeds 50.

# *6 Experimental Results*

Finally, in the last Section of this chapter, we present experimental results from testing the proposed approach in the domains of beach and mountain vacation images.

First, two individual domain ontologies were developed to represent the domain concepts of interest and their spatial relations. For the case of the beach vacation domain and under the current implementation, six concepts, namely: *Sky*, *Sea*, *Sand*, *Plant*, *Cliff* and *Person*, have been defined *a priori*. On the other hand, seven concepts, namely *Rock*, *Snow*, *Ground*, *Vegetation*, *Sky*, *Person* and *Water*, have been defined by domain experts for the case of the mountain vacation domain.

To acquire the visual descriptors prototypes and the membership values for the spatial relations, a training set of 200 images was assembled (100 for every domain) and manually annotated according to the domain ontology, using a variety of beach/mountain vacations images. Subsequently, a segmentation process was applied as described previously, and the *Dominant Colour*, *Homogeneous Texture* and *Region Shape* descriptors of the annotated segments were extracted. Approximately 10 prototype descriptor instances resulted for each of the defined domain concepts after the elimination of the redundant ones (i.e., of prototypes almost identical to each other that do not offer any additional discriminative power). Additionally, the degree to which each spatial relation is satisfied was estimated for each pair of segments and thus, following the already described procedure, the domain ontology spatial relations were enhanced with fuzzy degrees for each possible combination of the defined domain concepts.

After building the domain knowledge, semantic annotation of images can be performed following the proposed approach. For each of the examined images, the steps described in the low-level visual information processing section, i.e., segmentation, descriptors extraction and spatial relations extraction, are performed at first. Then, based on the prototype descriptor instances, initial hypotheses are generated for the examined image segments, following the matching approach described in section 3.1, which are in turn refined through the application of the context analysis presented in section 5. Finally, the updated graded hypotheses along with the extracted spatial relations are passed to the genetic algorithm, which is the one that determines the final image interpretation.

Quantitative performance measures are given in Tables I-II, in terms of precision and recall for the two examined domains. It must be noted that for the numerical

evaluation, any object present in the examined test set images that was not included in the domain ontologies was not taken into account. Indicative results are given in Fig. 10 and Fig. 11, showing the input image and the annotations resulting from the application of the genetic algorithm on the initial hypotheses and on the hypotheses refined by the context.

Visual context aids the overall labeling process, although in some concept cases we observe a marginal effect. An overall improvement of approximately 8.34% is given for the accuracy of the beach vacation domain after the final interpretation (Table I), whereas a 9.88% accuracy improvement is observed in the case of the mountain vacation domain (Table II), a fact mainly justified by the diversity and the quality of the provided image data set. Apart from that, the efficiency of the combination of two optimization steps (i.e., visual context together with a genetic algorithm) depends also heavily on the particularity of each specific concept; for instance, in Table I we observe that after the final interpretation of the images, the precision for the concept *Plant* improves with an overwhelming 248.03%, whereas in Table II, concept *Ground*'s precision and recall values are improved by 685.60% and 98.20%, respectively.

Adding visual context and a genetic algorithm to the semantic image analysis process is not an expensive process, in terms of computational complexity or timing. Average timing measurements for the overall process on the utilized dataset of images illustrate that it is a rather fast process. Based on our implementation, initial color image segmentation resulting to approximate 30-40 regions requires about 10 seconds, while visual descriptors extraction and initial region labeling are the major bottleneck, requiring 60 and 30 seconds, respectively. Comparing to the above numbers, all proposed algorithms (visual context and genetic algorithm) have significantly lower computational time, in the order of 1 second.

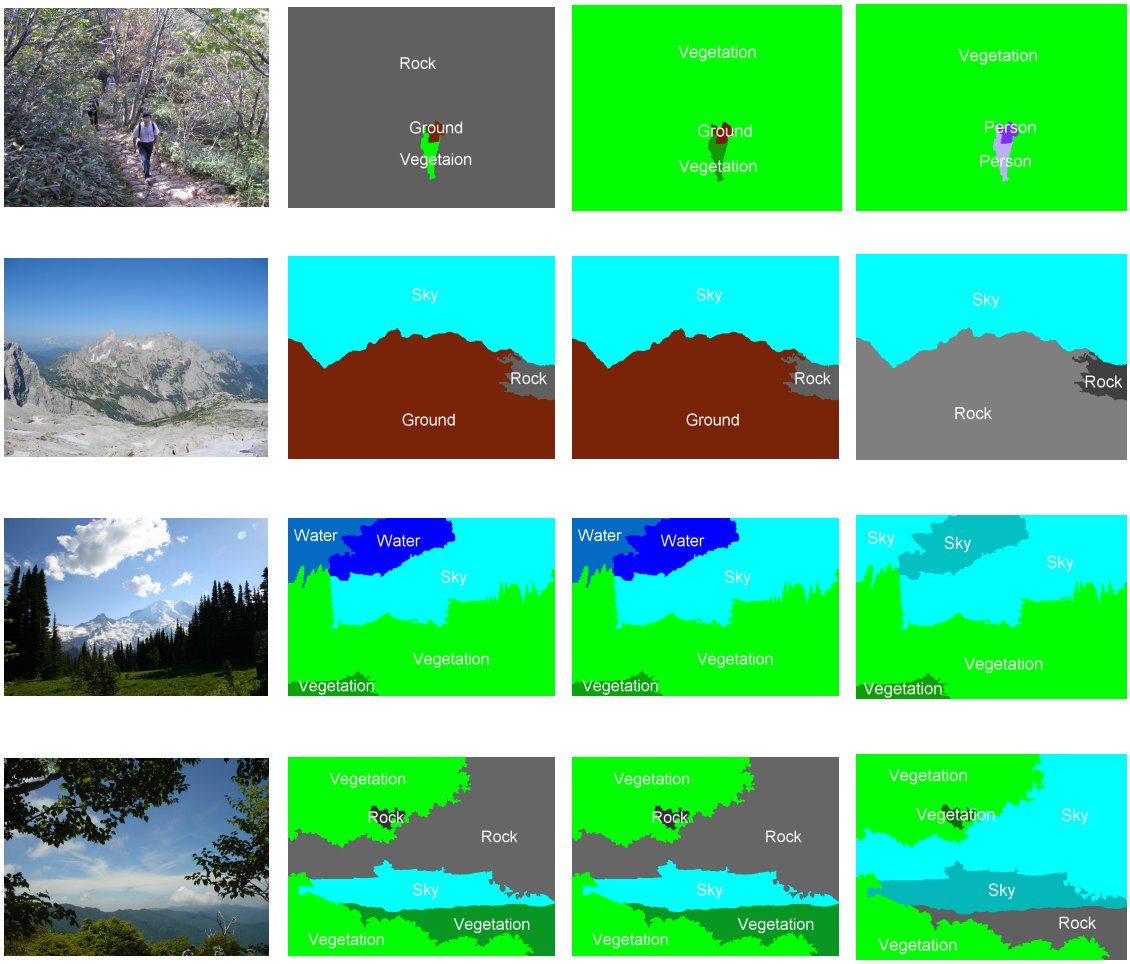
| Object | Initial hypothesis | | Hypothesis Refinement (visual context) | | Final Interpretation (genetic algorithm) | |
|---|---|---|---|---|---|---|
| | precision | recall | precision | recall | precision | recall |
| Sky | 83.33% | 94.74% | 92.78% | 94.74% | 95.79% | 92.86% |
| Sea | 93.55% | 87.00% | 90.95% | 95.50% | 94.50% | 90.00% |
| Cliff | 51.92% | 65.85% | 59.02% | 87.81% | 82.93% | 69.39% |
| Plant | 17.24% | 50.00% | 23.53% | 40.00% | 60.00% | 33.33% |
| Sand | 82.69% | 94.51% | 89.58% | 94.51% | 96.70% | 95.65% |
| Person | 97.03% | 71.02% | 98.99% | 71.02% | 81.16% | 99.12% |
| **Accuracy** | **82.76%** | | **87.07%** | | **89.66%** | |

Table I. Numerical evaluation for the beach vacation domain.

| Object | Initial Hypothesis | | Hypothesis Refinement (visual context) | | Final Interpretation (genetic algorithm) | |
|---|---|---|---|---|---|---|
| | precision | recall | precision | recall | precision | recall |
| Rock | 26.67% | 28.57% | 40.00% | 28.57% | 53.33% | 57.14% |
| Snow | 75.00% | 60.00% | 75.00% | 60.00% | 60.00% | 60.00% |
| Ground | 12.50% | 50.00% | 14.29% | 50.00% | 98.20% | 99.10% |
| Vegetation | 87.00% | 88.78% | 85.32% | 94.90% | 90.00% | 91.84% |
| Sky | 93.85% | 85.92% | 95.31% | 85.92% | 95.71% | 94.37% |
| Person | 37.50% | 33.33% | 33.33% | 22.22% | 50.00% | 55.56% |
| Water | 60.00% | 60.00% | 60.00% | 60.00% | 100.00% | 60.00% |
| **Accuracy** | **79.02%** | | **81.46%** | | **86.83%** | |

Table II. Numerical evaluation for the mountain vacation domain.


Fig. 10. Experimental results for the *beach* vacation domain – column (a) displays the input image, (b) the initial hypotheses, (c) the hypotheses refinement and (d) the final annotation of the image.




(a) (b) (c) (d)

Fig. 11. Experimental results for the *mountain* vacation domain - column (a) displays the input image, (b) the initial hypotheses, (c) the hypotheses refinement and (d) the final annotation of the image.

## 7 Conclusions

In this chapter, we presented our current research, view and implementation within the aceMedia approach to semantic image analysis. This is formulated as an optimization problem that couples ontologies with a genetic algorithm. The employed knowledge considers both high- and low-level information, represented using an ontology paradigm. The employed high-level knowledge includes the general domain knowledge in terms of concepts of interest and their spatial relations, as well as contextual knowledge in form of fuzzy ontological relations, whereas low-level knowledge consists of low-level visual descriptors required for the analysis process. Following such an approach, images from different domains can be semantically annotated, as long as the knowledge based is appropriately populated. The use of ontologies, due to the well-defined semantics that they provide, enables as well the application of inference services on top of the defined conceptualization that can lead to further enhanced annotations that can be inferred based on spatial reasoning. As illustrated within our experimentations, the proposed system achieves satisfactory results that are further improved through the exploitation of contextual knowledge. Thereby, the use of a genetic algorithm to treat image interpretation as an optimization problem is justified, as well as the added value entailed by the introduction and utilization of context into the analysis and interpretation chain.

## Acknowledgement

The work presented herein was partially supported by the European Commission under contract FP6-001765 aceMedia, FP6-027026 K-Space and FP6-507482 Knowledge-Web.

## References

[aceMedia] Integrating knowledge, semantics, and content, for user centered intelligent media services : the aceMedia project http://www.acemedia.org

Adamek, T., O'Connor, N., Murphy, N. (2005). Region-based Segmentation of Images Using Syntactic Visual Features. In Proc. of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Montreux, Switzerland.

Al-Khatib, W., Day, Y.F., Ghafoor, A., Berra, P.B. (1999). Semantic Modeling and Knowledge Representation in Multimedia Databases. IEEE Transactions on Knowledge and Data Engineering, 11(1).

Allen, J.F. (1983). Maintaining knowledge about temporal intervals. Communications of the ACM, 26(1):832–843.

Assfalg, J., Berlini, M., Del Bimbo, A., Nunziat, W., Pala, P. (2005). Soccer Highlights Detection and Recognition using HMMs. IEEE International Conference on Multimedia & Expo (ICME), 825-828.

Athanasiadis, Th., Tzouvaras, V., Petridis, K., Precioso, F., Avrithis, Y., & Kompatsiaris, I. (2005). Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content. In Proc. of SemAnnot '05, Galway, Ireland.

Benitez, A. B., Chang, S. F. (2003). Image Classification Using Multimedia Knowledge Networks. In Proc. IEEE Int. Conf. on Image Processing (ICIP03), Barcelona, Spain.

Benitez, A., Zhong, D., Chang, S. & Smith, J. (2001). MPEG-7 MDS Content Description Tools and Applications. In Proc. of International Conference on Computer Analysis of Images and Patterns (CAIP), Warsaw, Poland.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific American, 28(5), 34-43.

Bloehdorn, S., Petridis, K., Saathoff, C., Simou, N., Tzouvaras, V., Avrithis, Y., Handschuh, S., Kompatsiaris, I., Staab, S., Strintzis, M.G. (2005). Semantic Annotation of Images and Videos for Multimedia Analysis. In Proc. of 2nd European Semantic Web Conference, (ESWC 2005), Heraklion, Greece.

Cohn, A., Bennett, B., Gooday, J. M., & Gotts. N. M. (1997). Representing and Reasoning with Qualitative Spatial Relations about Regions, pages 97–134. Kluwer Academic Publishers.

Dasiopoulou, S., Mezaris, V., Kompatsiaris, I., Papastathis, V.K., Strintzis, M.G. (2005). Knowledge-Assisted Semantic Video Object Detection. IEEE Transactions, CSVT, Special Issue on Analysis and Understanding for Video Adaptation, 15(10), 1210–1224.

Edmonds, B. (1999). The Pragmatic Roots of Context. In Proc. of the 2nd International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-99), LNAI, vol. 1688, pp. 119-132, Berlin, Springer.

Gangemi, A., Guarino, N., Masolo, C. Oltramari, A., & Schneider, L. (2002). Sweetening ontologies with DOLCE, in Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, In Proc. of the 13th International Conference on Knowledge Acquisition, Modeling and Management, EKAW, LNCS, vol. 2473, Siguenza, Spain.

Goldberg, D., Deb, K. (1991). A comparative analysis of selection schemes used in genetic algorithms. In Foundations of Genetic Algorithms, G. Rawlins, 69–93.

Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specification. Knowledge Acquisition 5: 199-220.

Henderson, J. M., Hollingworth, A. (1999). High level scene perception. Annu. Rev. Psychol., vol. 50, pp. 243–271.

Hollink, L., Little, S., Hunter, J. (2005). Evaluating the Application of Semantic Inferencing Rules to Image Annotation. 3rd International Conference on Knowledge Capture (K-CAP05), Banff, Canada.

Hollink, L., Nguyen, G., Schreiber, G., Wielemaker, J., Wielinga, B., Worring, M. (2004). Adding Spatial Semantics to Image Annotations. In Proc. of International Workshop on Knowledge Markup and Semantic Annotation, ISWC.

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

ISO/IEC 15938-5 FCD Information Technology—Multimedia Content Description Interface— Part 5: Multimedia Description Schemes, March 2001, Singapore.

ISO/IEC FDIS 15938-5, ISO/IEC JTC 1/SC 29 M 4242, Information Technology Multimedia Content Description Interface Part 5: Multimedia Description Schemes, pp. 442-448, October 2001.

Klir, G., Yuan, B. (1995). Fuzzy Sets and Fuzzy Logic, Theory and Applications. New Jersey, Prentice Hall.

Mich, O., Brunelli, R., & Modena, C.M. (1999). A survey on video indexing. Journal of Visual Communications and Image Representation, 10:78–112.

Milanese, R. (1993). Detecting salient regions in an image: from biology to implementation. PhDThesis, University of Geneva, Switzerland.

Millet, C., Bloch, I., Hede, P., Moellic, P.-A. (2005). Using relative spatial relationships to improve individual region recognition. In Proc. of EWIMT, London.

Mitchell, T. (1999), Machine Learning and Data Mining, Communications of the ACM.

Miyamoto, S. (1990). Fuzzy Sets in Information Retrieval and Cluster Analysis. Kluwer Academic Publishers, Dordrecht / Boston / London.

MPEG-7 Visual Experimentation Model (XM) (2001), Version 10.0, ISO/IEC/JTC1/SC29/WG11, Doc. N4062.

Mylonas, Ph., Athanasiadis, Th., & Avrithis, Y. (2006). Improving image analysis using a contextual approach. In Proc. of Internationl Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Seoul, Korea.

Mylonas, Ph., & Avrithis, Y. (2005). Context modeling for multimedia analysis and use. In Proc. of 5th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT ), Paris, France.

Osberger, W., Maeder, A. J. (1998). Automatic Identification of Perceptually Important Regions in an Image. Proceedings of IEEE International Conference on Pattern Recognition.

Oliva, A., Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comp. Vis., vol. 42, pp. 145–175.

Papadias, D., & Theodoridis, Y. (1997). Spatial relations, minimum bounding rectangles, and spatial data structures. International Journal of Geographical Information Science, 11:111–138.

Papadopoulos, G. Th., Mezaris, V., Dasiopoulou, S., Kompatsiaris, I. (2006). Semantic Image Analysis Using a Learning Approach and Spatial Context. International Conference on Semantics and Digital Media Technologies (SAMT), Athens, Greece.

Rapantzikos, K., Avrithis, Y., Kollias, S. (2005). On the use of spatiotemporal visual attention for video classification. In Proc. of International Workshop on Very Low Bitrate Video Coding (VLBV '05), Sardinia, Italy.

Saathoff, C., Petridis, K., Anastasopoulos, D., Timmermann, N., Kompatsiaris I., & Staab, S. (2006). M-OntoMat-Annotizer: Linking Ontologies with Multimedia Low-Level Features for Automatic Image Annotation," Demos and Posters of the 3rd European Semantic Web Conference (ESWC), Budva, Montenegro.

Simou, N., Saathoff, C., Dasiopoulou, S., Spyrou, E., Voisine, N., Tzouvaras, V., Kompatsiaris, I., Avrithis, Y., Staab, S. (2005a). An Ontology Infrastructure for Multimedia Reasoning. International Workshop VLBV05, Sardinia, Italy.

Simou, N., Tzouvaras, V., Avrithis, Y., Stamou, G., & Kollias, S. (2005). A visual descriptor ontology for multimedia reasoning. In Proc. of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Montreux, Switzerland.

Skiadopoulos, S., Giannoukos, C., Sarkas, N., Vassiliadis, P., Sellis, T., Koubarakis, M. (2005). 2D topological and direction relations in the world of minimum bounding circles. IEEE Transactions on Knowledge and Data Engineering, 17(12), 1610-1623.

Skiadopoulos, S., & Koubarakis, M. (2004). Composing cardinal direction relations. Artificial Intelligence, 152:143–171.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R. (2000). Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 1349-1380.

Staab, S., Studer, R. (2004). Handbook on Ontologies, International Handbooks on Information Systems, Springer-Verlag, Heidelberg.

Stamou, G., & Kollias, S. (eds) (2005). Multimedia Content and the Semantic Web: Methods, Standards and Tools, John Wiley & Sons Ltd.

Tsechpenakis, G., Akrivas, G., Andreou, G., Stamou, G, Kollias, S. (2002). Knowledge-Assisted Video Analysis and Object Detection. In Proc. European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems (Eunite02), Algarve, Portugal.

Voisine, N., Dasiopoulou, S., Mezaris, V., Spyrou, E., Athanasiadis, Th., Kompatsiaris, I., Avrithis, Y., & Strintzis, M.G. (2005). Knowledge-assisted video analysis using a genetic algorithm. In Proc. of the 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005).

Wang, Y., Makedon, F., Ford, J., Shen, L., Golding, D. (2004). Generating Fuzzy Semantic Metadata Describing Spatial Relations from Images using the R-Histogram. JCDL , Tucson, Arizona, USA.

W3C, Semantic Web (2006). http://www.w3.org/2001/sw

W3C, RDF (2004). http://www.w3.org/RDF/

W3C, RDF Reification (2004). http://www.w3.org/TR/rdf-schema/

Zhang, L., Lin, F., Zhang, B. (2001). Support Vector Machine Learning For Image Retrieval. In Proc. of of International Conference on Image Processing, (2) 721-724.

Zhao, J., Shimazu, Y., Ohta, K., Hayasaka, R., Matsushita, Y. (1996). An Outstandingness Oriented Image Segmentation and its Applications. In Proc. of the International Symposium on Signal Processing and its Applications.

## *Author Bios*

**Giorgos Th. Papadopoulos,** was born in Thessaloniki, Greece in 1982. He received the Diploma degree in Electrical and Computer Engineering from Aristotle University of Thessaloniki, Greece in 2005. Currently he is pursuing his Ph.D. degree at the former University and he is a Postgraduate Research Fellow with the Informatics and Telematics Institute (ITI) / Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece. His research interests include still image segmentation, knowledge-assisted multimedia analysis, content-based and semantic multimedia indexing and retrieval, information extraction from multimedia, multimodal analysis and adaptive learning techniques. He is a member of the Technical Chamber of Greece.

**Phivos Mylonas**, MSc (Computer Science), is currently a Researcher by the Image, Video and Multimedia Laboratory. He obtained his Diploma in Electrical and Computer Engineering from the National Technical University of Athens in 2001, his Master of Science in Advanced Information Systems from the National & Kapodestrian University of Athens in 2003 and is currently pursuing his Ph.D. degree at the former University. His research interests lie in the areas of content-based information retrieval, visual context representation and analysis, knowledge-assisted multimedia analysis, issues related to personalization, user adaptation, user modeling and profiling.

**Dr. Vasileios Mezaris** received the Diploma degree and PhD in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2001 and 2005, respectively. He is a postdoctoral research fellow with the Informatics and Telematics Institute/Centre for Research and Technology Hellas, Thessaloniki, Greece. His research interests include image and video analysis, content-based and semantic image and video retrieval, ontologies, multimedia standards, knowledge-assisted multimedia analysis, knowledge extraction from multimedia, medical image analysis. He is a member of the IEEE and the Technical Chamber of Greece.

**Dr. Yannis Avrithis** was born in Athens, Greece in 1970. He received the Diploma degree in Electrical and Computer Engineering from the National Technical University of Athens in 1993, the M.Sc. degree in Communications and Signal Processing (with Distinction) from the Department of Electrical and Electronic Engineering of Imperial College of Science, Technology and Medicine, University of London, in 1994, and the Ph.D. degree in ECE from NTUA in 2001. He is currently a senior researcher at the Image, Video and Multimedia Systems Laboratory of NTUA. His research interests include spatiotemporal image/video segmentation, knowledge-assisted multimedia analysis and retrieval and personalization.

**Dr. Ioannis Kompatsiaris** received the Diploma degree in electrical engineering and the Ph.D. degree in 3-D model based image sequence coding from Aristotle University of Thessaloniki, Greece, in 1996 and 2001, respectively. He is a Senior Researcher with the Informatics and Telematics Institute. His research interests include semantic multimedia analysis, indexing and retrieval, multimedia and the Semantic Web, knowledge structures, reasoning and personalization for multimedia applications. He is the coauthor of 6 book chapters, 18 papers in refereed journals and more than 60 papers in international conferences. He is a member of IEEE and of the IEE VIE TAP.

## *Full contact details*

**Giorgos Th. Papadopoulos**
PhD Candidate
Multimedia Knowledge Group
Informatics and Telematics Institute
Centre for Research and Technology Hellas
1st Km. Thermi-Panorama Road
P.O. Box 60361, 57001 Thermi-Thessaloniki, Greece
Tel: +30 2310 464160 (ext. 125), Fax: +30 2310 464164
e-mail: papad@iti.gr
WWW: http://mkg.iti.gr

**Phivos Mylonas, M.Sc.**
Researcher
National Technical University of Athens
School of Electrical Engineering
Image, Video and Multimedia Laboratory
Iroon Polytechneioy 9, P.C. 157 80, Zographoy Campus, Athens, Greece
Electrical Engineering Building, Office 11.23, 1st Floor
Tel: +30 210 772 4351
Fax: +30 210 772 2492
e-mail: fmylonas@image.ntua.gr
WWW: http://www.image.ntua.gr/~fmylonas/

**Dr. Vasileios Mezaris**
Postdoctoral Research Fellow
Multimedia Knowledge Group
Informatics and Telematics Institute
Centre for Research and Technology Hellas
1st Km. Thermi-Panorama Road
P.O. Box 60361, 57001 Thermi-Thessaloniki, Greece
Tel: +30 2310 464160 (ext. 127), Fax: +30 2310 464164
e-mail: bmezaris@iti.gr
WWW: http://mkg.iti.gr

**Dr. Yannis Avrithis**.
Senior Researcher
National Technical University of Athens
School of Electrical Engineering
Image, Video and Multimedia Laboratory
Iroon Polytechneioy 9, P.C. 157 80, Zographoy Campus, Athens, Greece
Electrical Engineering Building, Office 11.23, 1st Floor
Tel: +30 210 772 4352
Fax: +30 210 772 2492
e-mail: iavr@image.ntua.gr
WWW: http://www.image.ntua.gr/~iavr/

**Dr. Yiannis Kompatsiaris**
Senior Researcher

Multimedia Knowledge Group
Informatics and Telematics Institute
Centre for Research and Technology Hellas
1st Km. Thermi-Panorama Road
P.O. Box 60361, 57001 Thermi-Thessaloniki, Greece
Tel: +30 2310 464160 (ext. 127), Fax: +30 2310 464164
e-mail: ikom@iti.gr
WWW: http://mkg.iti.gr