Semantic Image Segmentation and Object Labeling

Thanos Athanasiadis, *Student Member, IEEE*, Phivos Mylonas, *Student Member, IEEE*, Yannis Avrithis, *Member, IEEE*, and Stefanos Kollias, *Member, IEEE*

***Abstract*—In this paper, we present a framework for simultaneous image segmentation and object labeling leading to automatic image annotation. Focusing on semantic analysis of images, it contributes to knowledge-assisted multimedia analysis and bridging the gap between semantics and low level visual features. The proposed framework operates at semantic level using possible semantic labels, formally represented as fuzzy sets, to make decisions on handling image regions instead of visual features used traditionally. In order to stress its independence of a specific image segmentation approach we have modified two well known region growing algorithms, i.e., *watershed* and *recursive shortest spanning tree*, and compared them to their traditional counterparts. Additionally, a visual context representation and analysis approach is presented, blending global knowledge in interpreting each object locally. Contextual information is based on a novel semantic processing methodology, employing fuzzy algebra and ontological taxonomic knowledge representation. In this process, utilization of contextual knowledge re-adjusts labeling results of semantic region growing, by means of fine-tuning membership degrees of detected concepts. The performance of the overall methodology is evaluated on a real-life still image dataset from two popular domains.**

***Index Terms*—Fuzzy region labeling, semantic region growing, semantic segmentation, visual context.**

I. Introduction

AUTOMATIC segmentation of images is a very challenging task in computer vision and one of the most crucial steps toward image understanding. A variety of applications, such as object recognition, image annotation, image coding and image indexing, utilize at some point a segmentation algorithm and their performance depends highly on the quality of the latter. It was acknowledged that ages-long research has produced algorithms for automatic image and video segmentation [19], [12], [36], as well as structuring of multimedia content [16], [8]. Soon it was realized that image and video segmentation cannot overcome problems like (partial) occlusion, nonrigid motion, oversegmentation, among many others and reach the necessary level of efficiency, precision and robustness for the aforementioned application scenarios.

Research has been focused on the combination of *global* image features with *local* visual features to resolve regional ambiguities. In [24], a Bayesian network for integrating knowledge from low-level and midlevel features was used for indoor/outdoor classification of images. A multilabel classification is proposed in [14], similar to a fuzzy logic based approach, to describe natural scenes that usually belong to multiple semantic classes. Fuzzy classification was employed in [38] for region-based image classification. Learning of a Bayesian Network with classifiers as nodes, in [34], maps low level video representation to high level semantics for video indexing, filtering and retrieval. Knowledge on relative spatial positions between regions of the image is used in [28] as additional constrains, improving individual region recognition.

The notion of scene *context* is introduced in [32] as an extra source for both object detection and scene classification. An extension of the work presented in [24] was the additional use of temporal context, as given by the dependence between images captured within a short period of time [13]. Given the results of image analysis and segmentation, annotation can be either semi-automatic, like in [26], where a visual concept ontology guides the expert through the annotation process, or fully automatic, like the keyword assignment to regions in [23]. Recently, formal knowledge representation languages, like ontologies [39], [36] and description logics (DLs) [35] were used for construction of domain knowledge and reasoning for high level scene interpretation.

Comparatively to this effort, still, human vision perception outperforms state-of-the-art computer's segmentation algorithms. The main reason for this is that human vision is based also in high level prior knowledge about the semantic meaning of the objects that compose the image. Moreover, erroneous image segmentation leads to poor results in recognition of materials and objects, while at the same time, imperfections of global image classification are responsible for deficient segmentation. It is rather obvious that limitations of one prohibit the efficient operation of the other.

In this paper, we propose an algorithm that involves simultaneous segmentation and detection of simple objects, imitating partly the way that human vision works. An initial region labeling is performed based on matching a region's low-level descriptors with concepts stored in an ontological knowledge base; in this way, each region is associated to a fuzzy set of candidate labels. A merging process is performed based on new similarity measures and merging criteria that are defined at the semantic level with the use of fuzzy sets operations. Our approach can be applied to every region growing segmentation algorithm [1], [21], [10], [30] with the necessary modifications. Region growing algorithms start from an initial partition of the image and then an iteration of region merging begins, based on similarity criteria until the predefined termination criteria are met. We adjust appropriately these merging processes as well as the termination criteria.

Manuscript received May 1, 2006; revised October 26, 2006. This research was supported in part by the European Commission under Contracts FP6-001765 aceMedia and FP6-027026 K-SPACE and in part by the Greek Secretariat of Research and Technology (PENED Ontomedia 03EΔ475). This paper was recommended by Guest Editor E. Izquierdo.

The authors are with the National Technical University of Athens, 15773 Athens, Greece (e-mail: thanos@image.ntua.gr; fmylonas@image.ntua.gr; iavr@image.ntua.gr; stefanos@cs.ntua.gr).

Digital Object Identifier 10.1109/TCSVT.2007.890636

We also propose a context representation approach to use on top of semantic region growing. We introduce a methodology to improve the results of image segmentation, based on contextual information. A novel ontological representation for context is introduced, combining fuzzy theory and fuzzy algebra [22] with characteristics derived from the Semantic Web, like reification [42]. In this process, the membership degrees of labels assigned to regions derived by the semantic segmentation are re-estimated appropriately, according to a context-based membership degree readjustment algorithm. This algorithm utilizes ontological knowledge, in order to provide optimized membership degrees for the detected concepts of each region in the scene.

Our experiments employ contextual knowledge derived from two popular domains, i.e., *beach* and *motorsports*. It is worth noting, that the contextualization part, together with the initial region labeling step are domain-specific and require the domain of the images to be provided as input. The rest of the approach is domain-independent; however, because of the two previously identified parts, the overall result is that the domain of images needs to be specified *a priori*, in order to apply the proposed methodology.

The structure of this paper is as follows. Section II deals with the representation of knowledge used. More specifically, in Section II-A an ontology-based knowledge representation is described, used for visual context analysis, while in Section II-B a graph-based representation is introduced for the purpose of image analysis. Section III describes the process of extraction and matching of visual descriptors for initial region labeling. Continuing in Section IV, two approaches of semantic segmentation are presented, based on the output of Section III and the knowledge described in Section II-B. In Section V, we propose a visual context methodology used on top of semantic segmentation that improves the results of object labeling. Finally, in Section VI, experimental setup is discussed and both descriptive and overall results are presented, while Section VII concludes the paper.

II. Knowledge Representation

A. Ontology-Based Contextual Knowledge Representation

Among all possible knowledge representation formalisms, ontologies [20], [25], [40] present a number of advantages. In the context of this work, ontologies are suitable for expressing multimedia content semantics in a formal machine-processable representation that allows manual or automatic analysis and further processing of the extracted semantic descriptions. As an ontology is a formal specification of a shared understanding of a domain, this formal specification is usually carried out using a subclass hierarchy with relationships among the classes, where one can define complex class descriptions (e.g., in DL [6] or OWL [43]). Among all possible ways to describe ontologies one can be formalized as follows:

$$O = \{C, \{R_{pq}\}\}, \text{ where } R_{pq} : C \times C \rightarrow \{0, 1\}. \quad (1)$$

In (1), O is an ontology, C is the set of concepts described by the ontology, p and q are two concepts $p, q \in C$ and, R_{pq} is the semantic relation amongst these concepts. The proposed knowledge model is based on a set of concepts and relations between them, which form the basic elements towards semantic interpretation of the present research effort. In general, semantic relations describe specific kinds of links or relationships between any two concepts. In the crisp case, a semantic relation either relates $R_{pq} = 1$ or does not relate $R_{pq} = 0$ a pair of concepts p, q with each other. Although almost any type of relation may be included to construct such knowledge representation, the two categories commonly used are *taxonomic* (i.e., ordering) and *compatibility* (i.e., symmetric) relations. However, as extensively discussed in [3], compatibility relations fail to assist in the determination of the context and the use of ordering relations is necessary for such tasks. Thus, a first main challenge is the meaningful utilization of information contained in taxonomic relations for the task of context exploitation within semantic image segmentation and object labeling.

In addition, for a knowledge model to be highly descriptive, it must contain a large number of distinct and diverse relations among concepts. A major side effect of this approach is the fact that available information will then be scattered among them, making each one of them inadequate to describe a context in a meaningful way. Consequently, relations need to be combined to provide a view of the knowledge that suffices for context definition and estimation. In this work we utilize three types of relations, whose semantics are defined in the MPEG-7 standard [31], [7], namely the *specialization* relation Sp, the *part of* relation P and the *property* relation Pr.

A last point to consider when designing such a knowledge model is the fact that real-life data often differ from research data. Real-life information is in principal governed by uncertainty and fuzziness, thus its modeling is based on fuzzy relations. For the problem at hand, the above commonly encountered crisp relations can be modeled as fuzzy ordering relations and can be combined for the generation of a meaningful fuzzy taxonomic relation. Consequently, to tackle such complex types of relations we propose a "fuzzification" of the previous ontology definition, as follows:

$$O_F = \{C, \{r_{pq}\}\}, \text{ where } r_{pq} = F(R_{pq}) : C \times C \rightarrow [0, 1]. \quad (2)$$

In (2) O_F defines a "fuzzified" ontology, C is again the set of all possible concepts it describes and r_{pq} denotes a fuzzy relation amongst two concepts $p, q \in C$. In the fuzzy case, a fuzzy semantic relation relates a pair of concepts p, q with each other to a given degree of membership, i.e., the value of r_{pq} lies within the [0, 1] interval. More specifically, given a universe U a crisp set C is described by a membership function $\mu_C : U \rightarrow \{0, 1\}$ (as already observed in the crisp case for R_{pq}), whereas according to [22], a *fuzzy set* F on C is described by a membership function $\mu_F : C \rightarrow [0, 1]$. We may describe the fuzzy set F using the widely applied sum notation [29]:

$$F = \sum_{i=1}^{n} c_i / w_i = \{c_1/w_1, c_2/w_2, \ldots, c_n/w_n\}$$

where $n = |C|$ is the cardinality of set C and concept $c_i \in C$. The membership degree w_i describes the membership function $\mu_F(c_i)$, i.e., $w_i = \mu_F(c_i)$, or for the sake of simplicity,

$w_i = F(c_i)$. As in [22], a *fuzzy relation* on C is a function $r_{pq} : C \times C \rightarrow [0,1]$ and its *inverse* relation is defined as $r_{pq}^{-1} = r_{qp}$. Based on the relations r_{pq} and, for the purpose of image analysis, we construct the following relation T with use of the above set of fuzzy taxonomic relations: Sp, P, and Pr:

$$T = Tr^t(Sp \cup P^{-1} \cup \mathrm{Pr}^{-1}). \tag{3}$$

In the aforementioned relations, fuzziness has the following meaning. High values of $Sp(p,q)$ imply that the meaning of q approaches the meaning of p, in the sense that when an image is semantically related to q, then it is most probably related to p as well. On the other hand, as $Sp(p,q)$ decreases, the meaning of q becomes "narrower" than the meaning of p, in the sense that an image's relation to q will not imply a relation to p as well with a high probability, or to a high degree. Likewise, the degrees of the other two relations can also be interpreted as conditional probabilities or degrees of implied relevance. MPEG-7 MDS [31] contains all types of semantic relations, defined together with their inverses. Sometimes, the semantic interpretation of a relation is not meaningful whereas the inverse is. In our case, the relation *part* $P(p,q)$ is defined as: p part q if and only if q is *part of* p. For example, let p be New York and q Manhattan. It is obvious that the inverse relation *part of* P^{-1} is semantically meaningful, since Manhattan is *part of* New York. Similarly for the *property* relation Pr, its inverse is selected. On the other hand, following the definition of the *specialization* relation $Sp(p,q)$, p is *specialization* of q if and only if q is a specialization in meaning of p. For example, let p be a mammal and q a dog; $Sp(p,q)$ means that dog is a specialization of a mammal, which is exactly the semantic interpretation we wish to use (and not its inverse). Based on these roles and semantic interpretations of Sp, P and Pr, it is easy to see that (3) combines them in a straightforward and meaningful way, utilizing inverse functionality where it is semantically appropriate, i.e., where the meaning of one relation is semantically contradictory to the meaning of the rest on the same set of concepts. Finally, the transitive closure Tr^t is required in order for T to be taxonomic, as the union of transitive relations is not necessarily transitive, as discussed in [4].

Representation of our concept-centric contextual knowledge model follows the Resource Description Framework (RDF) standard [41] proposed in the context of the Semantic Web. RDF is the framework in which Semantic Web metadata statements are expressed and usually represented as graphs. The RDF model is based upon the idea of making statements about resources in the form of a subject-predicate-object expression. Predicates are traits or aspects about a resource and express a relationship between the subject and the object. Relation T can be visualized as a graph, in which every node represents a concept and each edge between two nodes constitutes a contextual relation between the respective concepts. Additionally each edge has an associated membership degree, which represents the fuzziness within the context model. Representing the graph in RDF is a straight forward task, since the RDF structure itself is based on a similar graph model.

The reification technique [42] was used in order to achieve the desired expressiveness and obtain the enhanced functionality introduced by fuzziness. Representing the membership degree associated with each relation is carried out by making a statement about the statement, which contains the degree information. Representing fuzziness with reified statements is a novel but acceptable way, since the reified statement should not be asserted automatically. For instance, having a statement such as "*Sky PartOf BeachScene*" and a membership degree of 0.75 for this statement, does obviously not entail, that sky is always a part of a beach scene.

```
<rdf:Description rdf:about="#s1">
    <rdf:subject rdf:resource="&dom;wrc"/>
    <rdf:predicate rdf:resource="&dom;specializationOf"/>
    <rdf:object>rdf:resource="&dom;rally"</rdf:object>
    <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Statement"/>
    <context:specializationOfrdf:datatype="http://www.w3.org/2001/XMLSchema#float">
    0.90</context:specializationOf>
</rdf:Description>
```

Fig. 1. Fuzzy relation representation: RDF reification.

A small clarifying example is provided in Fig. 1 for an instance of the specialization relation Sp. As already discussed, $Sp(x,y) > 0$ means that the meaning of x "includes" the meaning of y; the most common forms of specialization are subclassing, i.e., x is a generalization of y, and thematic categorization, i.e., x is the thematic category of y. In the example, the RDF subject *wrc* (World Rally Championship) has *specializationOf* as an RDF predicate and *rally* forms the RDF object. Additionally, the proposed reification process introduces a statement about the former statement on the *specializationOf* resource, by stating that 0.90 is the membership degree to this relation.

The ontology-based contextual knowledge model representation described in this section was used to construct the contextual knowledge for the series of experiments presented in Section VI. Two domain ontologies were 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 (i.e., in the *beach* domain: *sea*, *sand*, *sky*, *person*, *plant*, *cliff* and in the *motorsports* domain: *car*, *asphalt*, *gravel*, *grass*, *vegetation*, *sky*), as well as their interconnecting semantic relations, in terms of the membership degrees that correspond to each one of the three fuzzy semantic relations utilized, i.e., Sp, P, and Pr. As opposed to concepts themselves, which are manually defined by domain experts, the degrees of membership for each pair of concepts of interest, which are required for the construction of the relations during the ontology building process, are extracted using a "training set" of 120 images and probabilistic statistical information derived from it. The two application domains, i.e., *beach* and *motorsports*, were selected based on their popularity and the amount of multimedia content available.

B. Graph Representation of an Image

An image can be described as a structured set of individual objects, allowing thus a straightforward mapping to a graph structure. In this fashion, many image analysis problems can be considered as graph theory problems, inheriting the solid theoretical grounds of the latter. Attributed relation graph (ARG)

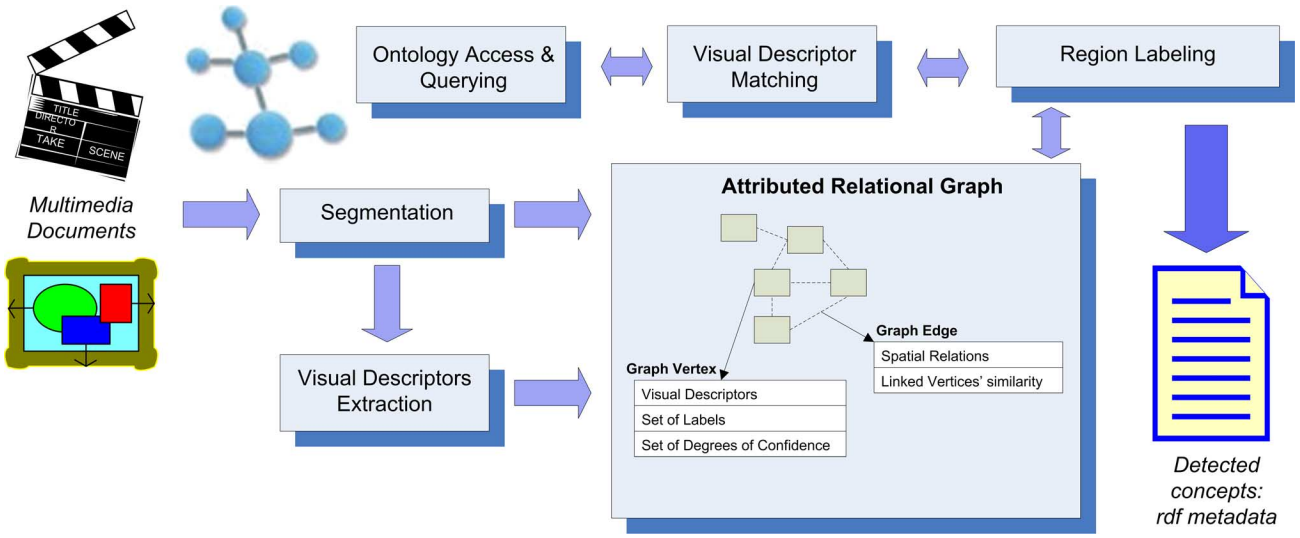

Fig. 2. Initial region labeling based on *ARG* and visual descriptor matching.

[9] is a type of graph often used in computer vision and image analysis for the representation of structured objects.

Formally, an ARG is defined by spatial entities represented as a set of vertices V and binary spatial relationships represented as a set of edges E : ARG $\equiv \langle V, E\rangle$. Letting G be the set of all connected, nonoverlapping regions/segments of an image, then a region $a \in G$ of the image is represented in the graph by vertex $v_a \in V$, where $v_a \equiv \langle a, D_a, L_a\rangle$. D_a is the ordered set of MPEG-7 visual descriptors characterizing the region in terms of low-level features, while $L_a = \sum_{i=1}^{|C|} c_i/\mu_a(c_i)$ is the fuzzy set of candidate labels for the region, extracted in a process described in the following Section. The adjacency relation between two neighbor regions $a, b \in G$ of the image is represented by graph's edge $e_{ab} \equiv \langle (v_a, v_b), s_{ab}\rangle \in E$. s_{ab} is a similarity value for the two adjacent regions represented by the pair (v_a, v_b). This value is calculated based on the semantic similarity of the two regions as described by the two fuzzy sets L_a and L_b

$$s_{ab} = \max_{c\in C}(\min(\mu_a(c), \mu_b(c))), a, b \in G. \tag{4}$$

The above formula states that the similarity of two regions is the default fuzzy union $(\max)$ over all common concepts of the default fuzzy intersection $(\min)$ of the degrees of membership $\mu_a(c)$ and $\mu_b(c)$ for the specific concept of the two regions a and b.

Finally, we consider two regions $a, b \in G$ to be connected when at least one pixel of one region is 4-connected to one pixel of the other. In an ARG, a neighborhood N_a of a vertex $v_a \in V$ is the set of vertices whose corresponding regions are connected to a: $N_a = \{v_b : e_{ab} \neq \emptyset\}, a, b \in G$. It is rather obvious now that the subset of ARG's edges that are incident to region a can be defined as: $E_a = \{e_{ab} : b \in N_a\} \subseteq E$.

The current approach (i.e., using two different graphs within this work) may look unusual to the reader at the first glance; however, using RDF to represent our knowledge model does not entail the use of RDF-based graphs for the representation of an image in the image analysis domain. Use of ARG is clearly favored for image representation and analysis purposes, whereas RDF-based knowledge model is ideal to store in and retrieve from a knowledge base. The common element of the two representations, which is the one that unifies and strengthens the current approach, is the utilization of a common fuzzy set notation, that bonds together both knowledge models. In the following Section we shall focus on the use of the ARG model and provide the guidelines for the fundamental initial region labeling of an image.

III. Initial Region Labeling

Our intention within this work is to operate on a semantic level where regions are linked to possible labels rather than only to their visual features. As a result, the above described ARG is used to store both the low level and the semantic information in a region-based fashion. Two MPEG-7 Visual Descriptors, namely *dominant color* (DC) and *homogeneous texture* (HT) [27], are used to represent each region in the low level feature-space, while fuzzy sets of candidate concepts are used to model high level information. For this purpose a knowledge assisted analysis algorithm, discussed in depth in [5], has been designed and implemented. The general architecture scheme is depicted in Fig. 2, where in the center lies the ARG, interacting with the rest processes.

The ARG is constructed based on an initial RSST segmentation [1] that produces a few tens of regions (approximately 30–40 in our experiments). For every region DC and HT are extracted (i.e., for region a: $D_a = [\mathrm{DC}_a\mathrm{HT}_a]$) and stored in the corresponding graph's vertex. Formal definition of the two descriptors as in [27] is

$$\mathrm{DC} \equiv [\{c_i, v_i, p_i\}, s], \quad i = 1, \ldots, N \tag{5}$$

where c_i is the ith dominant color, v_i the color's variance, p_i the color's percentage value, s the spatial coherency, and N can be

up to eight. The distance function for two descriptors $\mathrm{DC}_1, \mathrm{DC}_2$ is

$$d_{\mathrm{DC}}(\mathrm{DC}_1, \mathrm{DC}_2) = \sqrt{\sum_{i=1}^{N_1} p_{1i}^2 + \sum_{j=1}^{N_2} p_{2j}^2 - \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} 2a_{1i,2j}p_{1i}p_{2j}} \tag{6}$$

where $a_{1i,2j}$ is a similarity coefficient between two colors. Similarly for HT we have

$$\mathrm{HT} \equiv [\mathrm{avg}, \mathrm{std}, e_1, \ldots, e_{30}, d_1, \ldots, d_{30}] \tag{7}$$

where avg is the average intensity of the region, std is the standard deviation of the region's intensity, and e_i and d_i are the energy and the deviation for thirty ($i \in [1, \ldots, 30]$) frequency channels. A distance function is also defined

$$d_{\mathrm{HT}}(\mathrm{HT}_1, \mathrm{HT}_2) = \sum_{i=1}^{N_{\mathrm{HT}}=62} \left| \frac{\mathrm{HT}_1(i) - \mathrm{HT}_2(i)}{\sigma_i} \right| \tag{8}$$

where σ_i is a normalization value for each frequency channel. For the sake of simplicity and readability, we will use the following two distance notations equivalently: $d_{\mathrm{DC}}(\mathrm{DC}_a, \mathrm{DC}_b) \equiv d_{\mathrm{DC}}(a, b)$ (similarly for d_{HT}). This is also justified as we do not deal with abstract vectors but with image regions a and b represented by their visual descriptors.

Region labeling is based on a matching process between the visual descriptors stored in each vertex of the ARG and the corresponding visual descriptors of all concepts $c \in C$, stored in the form of prototype instances $P(c)$ in the ontological knowledge base. Matching of a region $a \in G$ with a prototype instance $p \in P(c)$ of a concept $c \in C$ is done by combining the individual distances of the two descriptors

$$\begin{aligned} d(a, p) &= d([\mathrm{DC}_a \mathrm{HT}_a], [\mathrm{DC}_p \mathrm{HT}_p]) \\ &= w_{\mathrm{DC}}(c) \cdot n_{\mathrm{DC}}(d_{\mathrm{DC}}(a, p)) \\ &\quad + w_{\mathrm{HT}}(c) \cdot n_{\mathrm{HT}}(d_{\mathrm{HT}}(a, p)) \end{aligned} \tag{9}$$

where d_{DC} and d_{HT} are given in (6) and (8), w_{DC} and w_{HT} are weights depending on each concept c and $w_{\mathrm{DC}}(c) + w_{\mathrm{HT}}(c) = 1, \forall \mathrm{c} \in \mathrm{C}$. Additionally, n_{DC} and n_{HT} are normalization functions and more specifically were selected to be linear

$$n(x) = \frac{x - d_{\min}}{d_{\max} - d_{\min}}, \quad n : [d_{\min} d_{\max}] \rightarrow [0\ 1] \tag{10}$$

where $d_{\min}$ and $d_{\max}$ are the minimum and maximum of the two distance functions d_{DC} and d_{HT}, respectively.

After exhaustive matching between regions and all prototype instances, the last step of the algorithm is to populate the fuzzy set L_a for all graph's vertices. The degree of membership of each concept c in the fuzzy set L_a is calculated as follows:

$$\mu_a(c) = 1 - \min_{p \in P(c)} d(a, p) \tag{11}$$

where $d(a, p)$ is given in (9). This process results to an initial fuzzy labeling of all regions with concepts from the knowledge base, or more formally to a set $L = \{L_a\}, a \in G$ whose elements are the fuzzy sets of all regions in the image.

This is obviously not a simple task and its efficiency depends highly on the domain where it is applied, as well as on the quality of the knowledge base. Main limitations of this approach are the dependency on the initial segmentation and the creation of representative prototype instances of the concepts. The latter is easier to be managed, whereas we deal with the former in this paper suggesting an extension based on region merging and segmentation on a semantic level.

IV. Semantic Region Growing

A. Overview

The major target of this work is to improve both image segmentation and labeling of materials and simple objects at the same time, with obvious benefits for problems in the area of image understanding. As mentioned in the introduction, the novelty of the proposed idea lies on blending well established segmentation techniques with midlevel features, like those we defined earlier in Section II-B.

In order to emphasize that this approach is independent of the selection of the segmentation algorithm, we examine two traditional segmentation techniques, belonging in the general category of region growing algorithms. The first is the watershed segmentation [10], while the second is the recursive shortest spanning tree (RSST) [30]. We modify these techniques to operate on the fuzzy sets stored in the ARG in a similar way as if they worked on low-level features (such as color, texture, *etc.*). Both variations follow in principles the algorithmic definition of their traditional counterparts, though several adjustments were considered necessary and were added. We call this overall approach semantic region growing (SRG).

B. Semantic Watershed

The watershed algorithm [10] owes its name to the way in which regions are segmented into catchment basins. A catchment basin is the set of points that is the local minimum of a height function (most often the gradient magnitude of the image). After locating these minima, the surrounding regions are incrementally flooded and the places where flood regions touch are the boundaries of the regions. Unfortunately, this strategy leads to oversegmentation of the image; therefore, a marker controlled segmentation approach is usually applied. Markers constrain the flooding process only inside their own catchment basin; hence the final number of regions is equal to the number of markers.

In our semantic approach of watershed segmentation, called semantic watershed, certain regions play the role of markers/seeds. During the construction of the ARG, every region $a \in G$ has been linked to a graph vertex $v_a \in V$ that contains a fuzzy set of labels L_a. A subset of all regions G are selected to be used as seeds for the initialization of the semantic watershed algorithm and form an initial set $S \subseteq G$. The criteria for selecting a region $s \in S$ to be a seed are as follows.

1) The height of its fuzzy set L_a (the largest degree of membership obtained by any element of L_a [22]) should be above a threshold: $h(L_a) > T_{\mathrm{seed}}$. Threshold T_{seed} is different for every image and its value depends on the distribution of all degrees of membership over all regions of the

particular image. The value of T_{seed} discriminates the top p percent of all degrees and this percentage p (calculated only once) is the optimal value (with respect to the objective evaluation criterion described in Section VI-A) derived from a training set of images.

2) The specific region has only one dominant concept, i.e., the rest concepts should have low degrees of membership comparatively to that of the dominant concept

$$h(L_a) > \sum_{c \in \{C - c*\}} \mu_a(c) \tag{12}$$

where $c*$ is the concetpt such that $\mu_a(c*) = h(L_a)$. These two constrains ensure that the specific region has been correctly selected as seed for the particular concept $c*$.

An iterative process begins checking every initial region-seed, $s \in S$, for all its direct neighbors N_s. Let $r \in N_s$ a neighbor region of s, or in other words, s is the propagator region of r: $s = p(r)$. We compare the fuzzy sets of those two regions $L_{p(r)}, L_r$ element by element and for every concept in common we measure the degree of membership of region r, for the particular concept $c, \mu_r(c)$. If it is above a merging threshold $\mu_r(c) > K^n \cdot T_{\text{merge}}$, then it is assumed that region r is semantically similar to its propagator and was incorrectly segmented and therefore, we merge those two. Parameter K is a constant slightly above one, which increases the threshold in every iteration n of the algorithm in a nonlinear way to the distance from the initial regions-seeds. Additionally region r is added in a new set of regions M_s^n (n denotes the iteration step, with $M_s^0 \triangleq s, M_s^1 \triangleq N_s$, etc.), from which the new seeds will be selected for the next iteration of the algorithm. After merging, the algorithm re-evaluates the degrees of membership of all concepts of L_r

$$\mu_{\hat{r}}(c) = \min(\mu_{p(r)}(c), \mu_r(c)) \tag{13}$$

where $p(r)$ is the propagator region of r.

The above procedure is repeated until the termination criterion of the algorithm is met, i.e., all sets of regions-seeds in step n are empty: $M_s^n = \emptyset$. At this point, we should underline that when neighbors of a region are examined, previous accessed regions are excluded, i.e., each region is reached only once and that is by the closest region-seed, as defined in the ARG.

After running this algorithm onto an image, some regions will be merged with one of the seeds, while other will stay unaffected. In order to deal with these regions as well, we run again the algorithm on a new ARG each time that consists of the regions that remained intact after all previous iterations. This hierarchical strategy needs no additional parameters, since every time new regions-seeds will be created automatically based on a new threshold T_{seed} (apparently with smaller value than before). Obviously, the regions created in the first pass of the algorithm have stronger confidence for their boundaries and their assigned concept than those created in a later pass. This is not a drawback of the algorithm; quite on the contrary, we consider this fuzzy outcome to be actually an advantage as we maintain all the available information.

C. Semantic RSST

Traditional RSST [30] is a bottom-up segmentation algorithm that begins from the pixel level and iteratively merges similar neighbor regions until certain termination criteria are satisfied. RSST is using internally a graph representation of image regions, like the ARG described in Section II-B. In the beginning, all edges of the graph are sorted according to a criterion, e.g., color dissimilarity of the two connected regions using Euclidean distance of the color components. The edge with the least weight is found and the two regions connected by that edge are merged. After each step, the merged region's attributes (e.g., region's mean color) is recalculated. Traditional RSST will also recalculate weights of related edges as well and resort them, so that in every step the edge with the least weight will be selected. This process goes on recursively until termination criteria are met. Such criteria may vary, but usually these are either the number of regions or a threshold on the distance.

Following the conventions and notation used so far, we introduce here a modified version of RSST, called Semantic RSST. In contrast to the approach described in the previous Section, in this case no initial seeds are necessary, but instead of this we need to define (dis)similarity and termination criteria. The criterion for ordering the edges is based on the similarity measure defined earlier in Section II-B. For an edge e_{ab} between two adjacent regions a and b we define its weight as follows:

$$w(e_{ab}) = 1 - s_{ab}. \tag{14}$$

Equation (14) can be expanded by substituting s_{ab} from (4). We considered that an edge's weight should represent the degree of dissimilarity between the two joined regions; therefore, we subtract the estimated value from one. Commutativity and associativity axioms of all fuzzy set operations (thus including default fuzzy union and default fuzzy intersection) ensure that the ordering of the arguments is indifferent. In this way all graph's edges are sorted by their weight.

Let us now examine in details one iteration of the semantic RSST algorithm. Firstly, the edge with the least weight is selected as: $e_{ab}^* = \text{argmin}_{e_{ab} \in E}(w(e_{ab}))$. Then regions a and b are merged to form a new region $\hat{a}$. Region b is removed completely from the ARG, whereas a is updated appropriately. This update procedure consists of the following two actions.

1) Update of the fuzzy set L_a by re-evaluating all degrees of membership in a weighted average fashion

$$\mu_{\hat{a}}(c) = \frac{A(a) \cdot \mu_a(c) + A(b) \cdot \mu_b(c)}{A(a) + A(b)}, \quad \forall c \in C. \tag{15}$$

The quantity $A(a)$ is a measure of the size (area) of region a and is the number of pixels belonging to this region.

2) Re-adjustment of the ARG's edges:
   a) Removal of edge e_{ab}.
   b) Re-evaluation of the weight of all affected edges e: the union of those incident to region a and of those incident to region b: $e \in E_a \cup E_b$.

This procedure continues until the edge e^* with the least weight in the ARG is above a threshold: $w(e^*) > T_w$. This

threshold is calculated in the beginning of the algorithm (similarly with the traditional RSST), based on the cumulative histogram of the weights of all edges E.

V. Visual Context

The idea behind the use of visual context information responds to the fact that not all human acts are relevant in all situations and this holds also when dealing with image analysis problems. Since visual context is a difficult notion to grasp and capture [33], we restrict it herein to the notion of ontological context. The latter is defined as part of the "fuzzified" version of traditional ontologies presented in Section II. In this section, the problems to be addressed include how to meaningfully readjust the membership degrees of the merged regions after the semantic region growing algorithm application and how to use visual context to influence the overall results of knowledge-assisted image analysis towards higher performance.

Based on the mathematical background described in detail in the previous subsections, we introduce the algorithm used to readjust the degree of membership $\mu_a(c)$ of each concept c in the fuzzy set L_a associated to a region $a \in G$ in a scene. Each specific concept $k \in C$ present in the application-domain's ontology is stored together with its relationship degrees r_{kl} to any other related concept $l \in C$. To tackle cases that more than one concept is related to multiple concepts, the term *context relevance* $\mathrm{cr}_{dm}(k)$ is introduced, which refers to the overall relevance of concept k to the *root element* characterizing each domain dm. For instance the *root element* of *beach* and *motorsports* domains are concepts beach and motorsports respectively. All possible routes in the graph are taken into consideration forming an exhaustive approach to the domain, with respect to the fact that all routes between concepts are reciprocal.

Estimation of each concept's value is derived from direct and indirect relationships of the concept with other concepts, using a meaningful *compatibility indicator* or distance metric. Depending on the nature of the domains under consideration, the best indicator could be selected using the *max* or the *min* operator, respectively. Of course the ideal distance metric for two concepts is again one that quantifies their semantic correlation. For the problem at hand and given the *beach* and *motorsports* domains, the *max* value is a meaningful measure of correlation for both of them. A simplified example, assuming that the only available concepts are motorsports (the *root element*—denoted as m), $\mathrm{asphalt}(a), \mathrm{grass}(g)$, and $\mathrm{car}(c)$ is presented in Fig. 3 and summarized in the following: let concept a be related to concepts m, g and c directly with: r_{am}, r_{ag}, and r_{ac}, while concept g is related to concept m with r_{gm} and concept c is related to concept m with r_{cm}. Additionally, c is related to g with r_{cg}. Then, we calculate the value for $\mathrm{cr}_{dm}(a)$

$$\mathrm{cr}_{dm}(a) = \max\{r_{am}, r_{ag}r_{gm}, r_{ac}r_{cm}, r_{ag}r_{cg}r_{cm}, r_{ac}r_{cg}r_{gm}\}. \tag{16}$$

The general structure of the degree of membership re-evaluation algorithm is as follows.

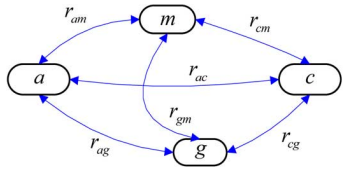

Fig. 3. Graph representation example. Compatibility indicator estimation.

1) Identify an optimal normalization parameter np to use within the algorithm's steps, according to the considered domain(s). The np is also referred to as domain similarity, or dissimilarity, measure and $np \in [0, 1]$.
2) For each concept k in the fuzzy set L_a associated to a region $a \in G$ in a scene with a degree of membership $\mu_a(k)$, obtain the particular contextual information in the form of its relations to the set of any other concepts: $\{r_{kl} : l \in C, l \neq k\}$.
   Calculate the new degree of membership $\mu_a(k)$ associated to region a, based on np and the context's relevance value. In the case of multiple concept relations in the ontology, relating concept k to more than one concepts, rather than relating k solely to the "root element" r^e, an intermediate aggregation step should be applied for k: $\mathrm{cr}_k = \max\{r_{kr^e}, \ldots, r_{\mathrm{km}}\}$. We express the calculation of $\mu_a(k)$ with the recursive formula

$$\mu_a^n(k) = \mu_a^{n-1}(k) - np(\mu_a^{n-1}(k) - \mathrm{cr}_k) \tag{17}$$

   where n denotes the iteration used. Equivalently, for an arbitrary iteration n

$$\mu_a^n(k) = (1 - np)^n \cdot \mu_a^0(k) + (1 - (1 - np)^n) \cdot \mathrm{cr}_k \tag{18}$$

   where $\mu_a^0(k)$ represents the original degree of membership.

In praxis, typical values for n reside between 3 and 5. Interpretation of both (17) and (18) implies that the proposed contextual approach will favor confident degrees of membership for a region's concept in conjunction to nonconfident or misleading degrees of membership. It will amplify their differences, while on the other hand it will diminish confidence in clearly misleading concepts for a specific region. Further, based on the supplied ontological knowledge it will clarify and solve ambiguities in cases of similar concepts or difficult-to-analyze regions.

Key point in this approach remains the definition of a meaningful normalization parameter np. When re-evaluating these values, the ideal np is always defined with respect to the particular domain of knowledge and is the one that quantifies their semantic correlation to the domain. In this work we conducted a series of experiments on a training set of 120 images for both application domains and selected the np that resulted in the best overall evaluation score values for each domain.

The proposed algorithm readjusts in a meaningful manner the initial degrees of membership, utilizing semantics in the form of the contextual information residing in the constructed "fuzzified" ontology. In the following Section we discuss the experimental setup of this work and present both descriptive and overall results.

VI. Experimental Results

A. Experiments Setup and Evaluation Procedure

In order to evaluate our work, we carried out experiments in the domains of *beach* and *motorsports*, utilizing a data set of 602 images in total, i.e., 443 *beach* and 159 *motorsports* images acquired either from the Internet or from personal collections. In the process of evaluating this work and testing its tolerance to imprecision in initial labels, we conducted a series of experiments with a subset of 482 images originating from the above data set, since 120 images (a 20% subset) was used as a training set for optimum parameter and threshold estimation, such as *np* and T_{seed}. It is a common fact [15], [17] that the most objective segmentation evaluation includes a relative evaluation method that employs a corresponding ground truth. For this purpose we developed an annotation tool for the manual construction of the ground truth. Human experts spent an effort to select and annotate the subset of images utilized during the evaluation steps. In order to demonstrate the proposed methodologies and keep track of each individual algorithm results, we integrated the described techniques into a single application enhanced with a graphical user interface.

The evaluation procedure is always particularly critical because it quantifies the efficiency of an algorithm, assisting scientific and coherent conclusions to be drawn. Since fuzzy sets were used throughout this work, we adopted fuzzy sets operations to evaluate the results. The final output of both semantic region growing variations, as well as context algorithm is a segmentation mask together with a fuzzy set L_a for any region a that contains all candidate object/region labels with their degrees of membership. The ground truth of an arbitrary image consists of a number of connected, nonoverlapping segments seg_i that are associated (manually) to a unique label.

First we calculate the overlap of each region a with each segment seg_i of the ground truth, as illustrated in the following equation, where the quantity $A(a)$ is again a measure of the size of a region a and is defined right after (15) in Section IV-C

$$\text{overlap}(a, \text{seg}_i) = \frac{A(a \cap \text{seg}_i)}{A(a)}. \tag{19}$$

Then we calculate the Dombi t-norm [18] with parameter $p = 3$ of the overlap and the membership degree of the corresponding (to the ground truth's segment) label

$$\text{score}_a(c) = \text{T}_{\text{Dombi}}(\text{overlap}(a, \text{seg}_i), \mu_a(c)) \tag{20}$$

where c is the concept characterizing seg_i. Doing this for L_a we calculate the total score of region a using Dombi t-conorm over all concepts

$$\text{score}_a = \perp_{\text{Dombi}}(\{\text{score}_a(c) : c \in L_a\}). \tag{21}$$

Due to associativity axiom of fuzzy sets t-conorms, (21) arguments' order is totally indifferent. The equation gives us an evaluation score of a particular region, which is not completely useless, nevertheless is not a measure for the whole image. This global measure is acquired by applying an aggregation operation on all individual score_a w.r.t. the size of each region

$$\text{score} = \frac{\sum_{a \in G} (A(a) \cdot \text{score}_a)}{\sum_{a \in G} A(a)}. \tag{22}$$

Equation (22) provides an overall performance evaluation score suitable for the herein presented algorithms against some ground truth.

The above evaluation strategy is also followed for assessing the segmentation results of the traditional watershed and RSST algorithms, which is necessary for comparison purposes with the proposed semantic approach. Obviously both traditional algorithms lack the semantic information (i.e., the fuzzy set L_a for every region), therefore, we need to insert this at the end of the process in order to be able to calculate the evaluation score. This is done by following exactly the methodology presented in Section III, i.e., extraction and matching of visual descriptors. The apparent difference with the semantic segmentation approach is that labels and degrees are not taken into consideration during segmentation but only at the end of the process.

B. Indicative Results and Discussion

The overall outcome from evaluation tests conducted on the entire dataset of images are promising, even in cases where detection of specific object labels is rather difficult; region growing is guided correctly and ends up in a considerable improvement of the traditional counterpart segmentation algorithm. Additionally final regions are associated to membership degrees providing a meaningful measure of belief for the segmentation accuracy.

In the following, we present three detailed sets of descriptive experimental results in order to illustrate the proposed techniques, implicating two images derived from the beach domain and one image from the motorsports domain. We also include an illustration of the semantic similarity between two adjacent regions, as well as a contextualization example on two arbitrary images to stress both our region similarity calculation approach and the aid of context in the process. To provide an assessment overview over both application domains, we present evaluation results over the entire dataset utilized, implementing both image segmentation algorithms with and without context optimization.

Each one of the descriptive image sets includes four images: (a) the original image; (b) the result of the traditional RSST; (c) of the semantic watershed; and (d) of the semantic RSST. The ground truth corresponding to the three image sets is included in Fig. 4. In the case of the traditional RSST, we predefined the final number of regions to be produced to be equal to that produced by the semantic watershed; in this fashion, segmentation results are better comparable.

Fig. 5 illustrates the first example derived from the *beach* domain. An obvious observation is that RSST segmentation performance in Fig. 5(b) is rather poor, given the specific image content; persons are merged with sand, whereas sea on the left and in the middle under the big cliff is divided into several regions, while adjacent regions of the same cliff are classified as different ones. The results of the application of the semantic

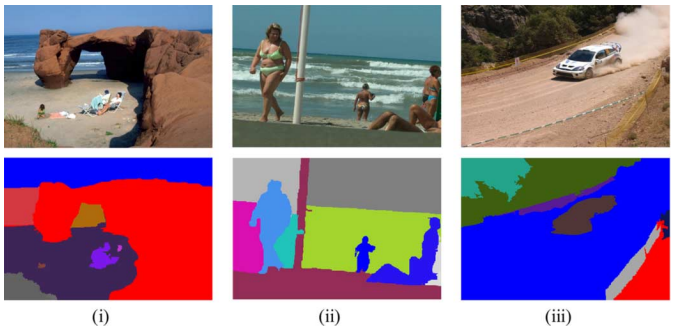

Fig. 4. Ground truth of two *beach* and one *motorsports* images.

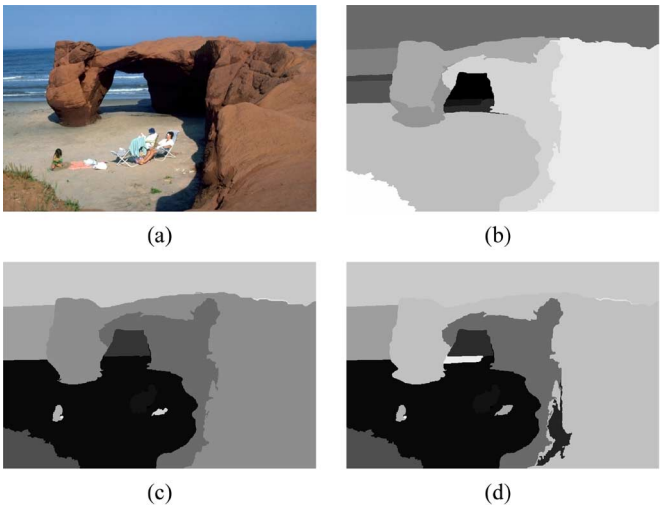

Fig. 5. Experimental results for the *beach* domain—Example 1. (a) Input image. (b) RSST segmentation. (c) Semantic watershed. (d) Semantic RSST.

watershed algorithm are shown in Fig. 5(c) and are considerably better. More specifically, we observe that both parts of sea are merged together and sea appears unified, both in the upper left corner and in the middle under the rock. The rocks on the right side are comprised of only two large regions, although their original variations in texture and color information. Splitting of the rocks into two regions is acceptable in this case, since one region comprises the thick shadow of the cliff, which typically confuses such a process.

Moreover, identification of the regions that constitute persons lying on the sand is sufficient, given the difficulty of the task, i.e., the fact that irrelevant objects are present in the foreground and that different low level variations insert a degree of uncertainty in the process. Good results are obtained also in the case of the bright shadow in the middle of the image, i.e., underneath the big cliff curve and below the corresponding sea part. This region is correctly identified as sand in contradiction to the dark shadow merged previously with the cliff. Finally, Fig. 5(d) illustrates the results of the application of our second semantic region growing approach, based on semantic RSST. In comparison to semantic watershed results, we observe small differences. For instance, the cliff on the right side is comprised by three regions and sea underneath the big cliff curve is also divided. Such variations in the results are expected, because of the nature of the semantic

TABLE I
DEGREES OF MEMBERSHIP OF EACH CONCEPT FOR FOUR NEIGHBORING REGIONS OF THE IMAGE OF FIG. 5(a)

Region	Concepts					
	Sky	Sea	Cliff	Plant	Sand	Person
a	0.66	0.82	0.67	0.65	0.68	0.75
b	0.74	0.77	0.64	0.64	0.69	0.73
c	0.68	0.79	0.68	0.65	0.75	0.67
d	0.90	0.77	0.67	0.65	0.64	0.98

TABLE II
SIMILARITY AND WEIGHTS OF THE EDGES BETWEEN THE FOUR NEIGHBORING REGIONS OF TABLE I

Edges	Concept						Sim/ty	Weight
	Sky	Sea	Cliff	Plant	Sand	Person	*s*	*w*
e_{ab}	0.66	0.77	0.64	0.64	0.68	0.73	**0.77**	**0.23**
e_{ac}	0.66	0.79	0.67	0.65	0.68	0.67	**0.79**	**0.21**
e_{ad}	0.66	0.77	0.67	0.65	0.64	0.75	**0.77**	**0.23**
e_{cd}	0.68	0.77	0.64	0.64	0.69	0.67	**0.77**	**0.23**

RSST algorithm, i.e., the latter is focused more on material detection. Overall quantitative results are following the described guidelines and their performance is good, given their individual total scores, as defined in previous subsection: 0.61 for RSST, 0.85 for semantic watershed and 0.78 for semantic RSST.

At this point, let us examine in detail a specific part of the image for one iteration of the semantic RSST algorithm. Initial segmentation and region labeling produced thirty regions in total with their associated labels. According to the ground truth, four of them (regions *a, b, c, d*) correspond to only one region, which is a sea region. These four regions form a subgraph G_{sea} of the ARG : $G_{\text{sea}} \equiv \langle V_{\text{sea}}, E_{\text{sea}} \rangle$, where $V_{\text{sea}} = \{v_a, v_b, v_c, v_d\}$ and $E_{\text{sea}} = \{e_{ab}, e_{ac}, e_{ad}, e_{cd}\}$.

In Table I, we illustrate the degrees of membership of each concept for those four regions. Based on these values and on (4) and (14), we calculate the similarity s of all neighbor regions and the weights of the corresponding edges, as illustrated in Table II. Utilizing the degrees obtained from Table I, we calculate the similarity of two regions for each concept, as depicted in the inner columns of Table II.

We can see that the edge with the least weight is e_{ac}, where $w(e_{ac}) = 0.21$ and that regions a, c have the greater similarity value: $s_{ac} = 0.79$, based on their common concept *sea*. Those two regions are merged and form a new region $\hat{a}$. According to (13) the fuzzy set of concepts for the new region is updated

$$\mu_{\hat{a}}(\text{sea}) = \frac{A(a) \cdot \mu_a(\text{sea}) + A(c) \cdot \mu_c(\text{sea})}{A(a) + A(c)}$$

and by substituting the values

$$\mu_{\hat{a}}(\text{sea}) = \frac{36054 \cdot 0.82 + 16011 \cdot 0.79}{36054 + 16011} \approx 0.811.$$

Similarly, we calculate $\mu_{\hat{a}}(\text{sky}), \mu_{\hat{a}}(\text{cliff})$, etc. Following the second step of the algorithm, edge e_{ac} is removed from the graph

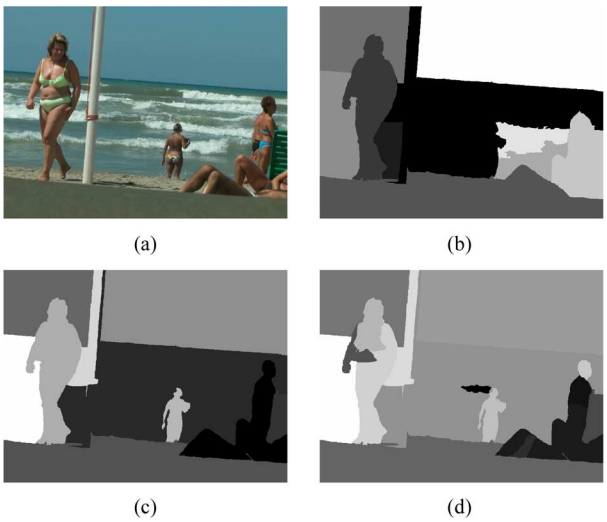

Fig. 6. Experimental results for the *beach* domain—Example 2. (a) Input image. (b) RSST segmentation. (c) Semantic watershed. (d) Semantic RSST.

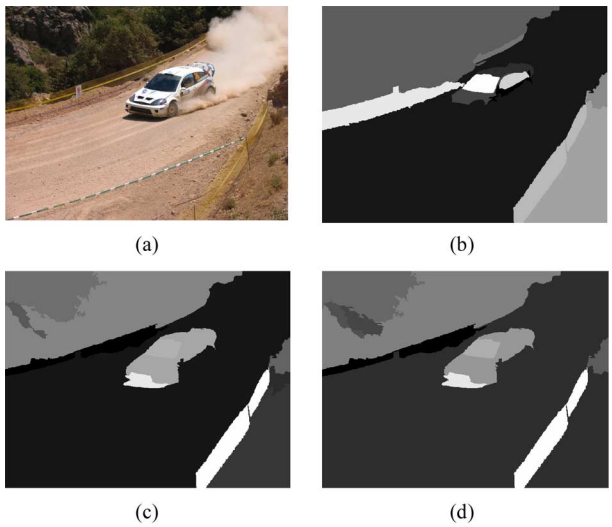

Fig. 7. Experimental results for the motorsports domain. (a) Input image. (b) RSST segmentation. (c) Semantic watershed. (d) Semantic RSST.

and all affected weights are recalculated according to (13) and the new fuzzy set $L_{\hat{a}}$.

In Fig. 6, RSST segmentation results [Fig. 6(b)] are again insufficient: some persons are unified with sea segments, while others are not detected at all and most sea regions are divided because of the waves. Semantic watershed application results into significant improvements [Fig. 6(c)]. Sea regions on the left part of the image are successfully merged together, the woman on the left is correctly identified as one region, despite the existence of variations in low level characteristics, i.e., green swimsuit versus color of the skin, etc. Persons on the right side are identified and not merged with sea or sand regions, having as a side effect the fact that there are multiple persons in the image and not just a single one. Very good results are obtained in the case of the sea in the right region, although it is inhomogeneous in terms of color and material because of the waving. We observe that it is successfully merged into one region and the person standing in the foreground is also identified as a whole. Finally, the semantic RSST algorithm in Fig. 6(d) performs similarly well. Small differences between semantic watershed and semantic RSST are justified by the fact that with the semantic RSST approach focus is given on material and not in objects in the image. Consequently, persons are identified with greater accuracy in the image and are segmented, but not wrongly merged, e.g., the woman on the left is composed by multiple regions due to the nature of the material or people on the right are composed by different regions. In terms of the objective evaluation score, results are verifying previous observations, namely RSST has a score of 0.82, semantic watershed a score of 0.90, and semantic RSST a score of 0.88.

Results from the *motorsports* domain are described in Fig. 7. More specifically, in Fig. 7(a) we present the original image derived from the World Rally Championship. Plain segmentation results [Fig. 7(b)] are again poor, since they do not identify correctly materials and objects in the image and incorrectly unify large portions of the latter into a single region. Fig. 7(c) and (d) illustrate distinctions between vegetation and cliff regions in the upper left corner of the image. Even different vegetation areas are identified as different regions in the same area. Furthermore, the car's windshield remains correctly a standalone region, because of its large color and material diversities in comparison to the regions in its neighborhood. Because of the difficulties and obstacles set by the nature of the image, the thick shadow in the front of the car is inevitably unified with the front dark part of the latter and the "gravel smoke" on the side is recognized as gravel, resulting into a deformation of the vehicle's chassis. These are two cases where both semantic region growing algorithms seem to perform poorly. This is due to the fact that the corresponding segments differ visually and the possible detected object is a composite one—in contradiction to the so far encountered material objects—and is composed by regions of completely different characteristics. Furthermore, on the right side of the image, the yellow ribbon is dividing two similar but not identical gravel regions, fact that is correctly identified by our algorithm. The main difference between the semantic watershed and semantic RSST approaches is summarized in the way they handle vegetation in the upper left corner of the image, with semantic RSST performing closer to the ground truth, since it detects the variations in vegetation and grass successfully. Finally, in terms of evaluation, we observe the following scores: 0.64 for RSST, 0.70 for semantic watershed and 0.71 for semantic RSST.

At this point we continue by presenting a detailed visualization of the contextualization step implemented within our approach. In general, our context algorithm successfully aids in the determination of regions in the image and corrects misleading behaviors, originating from over- or under-segmentation, by meaningfully adjusting their membership degrees. Utilizing the training set of 120 images, we selected the *np* value that resulted in the best overall evaluation score values for each domain. In other words, one *np* value is used for images belonging to the *beach* domain, namely $np = 0.15$ and a different one is utilized when dealing with images from the *motorsports* domain, i.e., $np = 0.20$ in this case.

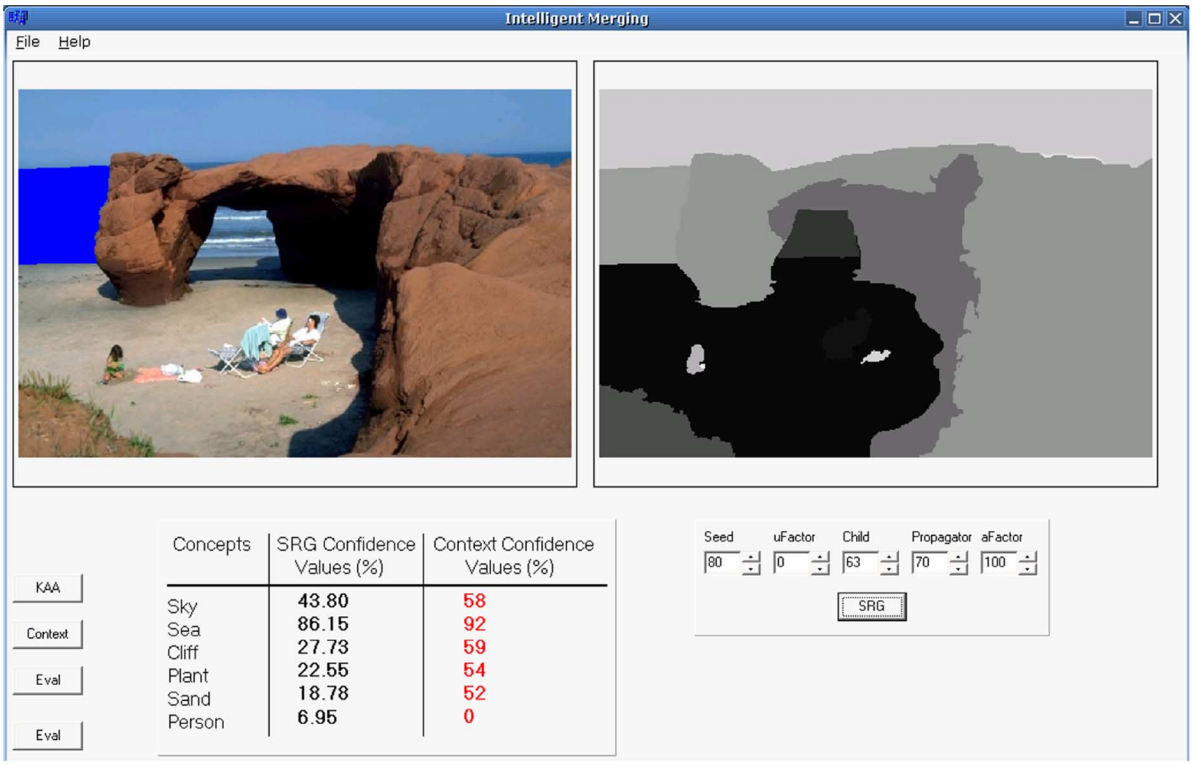

Fig. 8. Contextual experimental results for the first *beach* image example.

In Fig. 8 we observe the contextualization step for the first *beach* image, presented within the developed contextual analysis tool. Contextualization, which works on a per region basis, is applied after semantic region growing, in order for its results to be meaningful. We have selected the unified sea region in the upper left part of the image, as illustrated by its artificial electric-blue color. The contextualized results are presented in red in the right column at the bottom of the tool. Context favors strongly the fact that the merged region belongs to sea, increasing its degree of membership from 86.15% to a crisp 92%.

The totally irrelevant (for the region) membership degree for person is extinguished, whereas medium degrees of membership for the rest of the possible *beach* concepts are slightly increased, due to the ontological knowledge relations that exist in the considered knowledge model. In all cases context normalizes results in a meaningful manner, i.e., the dominant concept is detected with increased degree of membership.

To illustrate further the aid of context in our approach, we also present Table III, which illustrates the merged region concepts together with their membership degrees *before* and *after* the aid of visual context in the case of the second *beach* image. Table III is provided in order to summarize the influence of context on the merged regions, indicating significant improvements in most cases. The first column of the Table represents the final merged region id after the application of our semantic image segmentation approach. Each of the next six concept columns includes a twofold value, i.e., the membership degree without and with the aid of context. Pairs of values in boldface indicate the ground truth for the specific region.

It is easy to observe that in the majority of cases context optimizes the final labeling results, in terms of improving the concept's membership degree. Ground truth values are highlighted in order to provide comparative results to the reader, since these are the values of interest during the evaluation process. For instance, when considering region 0, which is a *sea* region according to the ground truth, context improves sea's membership degree by an 11.11% increase from 0.81 to 0.90. Similarly, considering region 18, context denotes a 13.10% increase regarding the actual *sky* concept, whereas region 29 illustrates a 6.69% increase of the membership degree for context, when tackling the *sand* concept. The above ground truth concept improvements (e.g., an overall average value of 12.55% for the concept *sea* and 6.17% for the concept *person*) are important, as depicted by their percentage increase and as they provide a basic evaluation tool of the proposed approach. As an overall conclusion, it is evident that a clear trend exists in most cases, i.e., the application of visual context affects positively the semantic segmentation process and this can be verified by the available ground truth information.

C. Overall Results

Finally, in the process of evaluating this work and testing its tolerance to imprecision in initial labels, we provide an evaluation overview of the application of the proposed methodology on the dataset of the *beach* domain. It must be pointed out that for the experiments regarding semantic watershed, the same value of T_{seed} was used. For the estimation of the value T_{seed}, we need to calculate the percentage p, as mentioned in Section IV-B. This is achieved by running the semantic watershed algorithm on the training set defined in Section VI-A and calculating the overall evaluation score for eight different values

TABLE III
FINAL DEGREES OF MEMBERSHIP *BEFORE* AND *AFTER* APPLYING VISUAL CONTEXT TO THE SECOND BEACH IMAGE—BOLDFACE INDICATE GROUND TRUTH INFORMATION

Region ID	Concepts											
	Sky		Sea		Cliff		Plant		Sand		Person	
	before	after	before	after	before	after	before	after	before	after	before	after
Region 0	0.66	0.72	***0.82***	***0.90***	0.67	0.73	0.65	0.69	0.68	0.70	0.75	0.74
Region 3	0.73	0.75	0.72	0.85	0.64	0.72	0.66	0.69	0.64	0.69	***0.73***	***0.75***
Region 7	0.69	0.73	***0.78***	***0.88***	0.66	0.72	0.65	0.69	0.72	0.72	0.93	0.81
Region 9	0.69	0.73	***0.66***	***0.83***	0.64	0.72	0.67	0.70	0.66	0.69	0.65	0.69
Region 12	0.66	0.71	0.70	0.85	0.67	0.73	0.65	0.69	0.74	0.73	***0.81***	***0.91***
Region 13	0.64	0.71	0.71	0.85	0.69	0.74	0.64	0.69	***0.73***	***0.75***	0.75	0.73
Region 15	0.90	0.82	***0.77***	***0.88***	0.67	0.73	0.65	0.69	0.64	0.69	0.98	0.84
Region 16	0.79	0.74	0.80	0.89	0.64	0.72	0.66	0.69	0.65	0.69	***0.73***	***0.75***
Region 18	***0.84***	***0.95***	0.83	0.90	0.67	0.73	0.89	0.80	0.65	0.69	0.00	0.40
Region 26	0.69	0.73	***0.90***	***0.93***	0.69	0.74	0.66	0.69	0.65	0.69	0.84	0.77
Region 29	0.76	0.76	0.61	0.71	0.68	0.74	0.74	0.73	***0.70***	***0.75***	0.87	0.79

TABLE IV
EVALUATION SCORES OF SEMANTIC WATERSHED ALGORITHM FOR THE TRAINING SET, WITH RESPECT TO PERCENTAGE p

***p* (%)**	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5
score	0.72	0.71	0.75	0.77	0.69	0.65	0.68	0.69

TABLE V
OVERALL AND PER CONCEPT DETECTION SCORES FOR THE ENTIRE BEACH DOMAIN

Concepts	*W*	*SW*	*SW+C*	*R*	*SR*	*SR+C*
Sky	0.93	0.95	0.94	0.90	0.93	0.93
Sea	0.79	0.80	0.84	0.81	0.83	0.89
Sand	0.72	0.79	0.81	0.72	0.82	0.88
Person	0.44	0.56	0.60	0.48	0.48	0.55
Cliff	0.61	0.75	0.77	0.66	0.76	0.78
Plant	0.67	0.70	0.71	0.63	0.68	0.69
Total	**0.69**	**0.77**	**0.79**	**0.70**	**0.77**	**0.78**

of p (see Table IV). In this way we acquired the best results for $p = 5.5\%$.

In Table V, detection results for each concept, as well as the overall score derived from the six beach domain concepts are illustrated. Scores are presented for six different algorithms: traditional watershed (W), semantic watershed (SW), semantic watershed with context (SW+C), traditional RSST (R), semantic RSST (SR), and semantic RSST with context (SR+C). Apparently, concept sky has the best score among the rest, since its color and texture are relatively invariable. Visual context indeed aids the labeling process, even with the overall marginal improvement of approximately 2%, given in Table V, a fact mainly justified by the diversity and the quality of the provided image data set. Apart from that, the efficiency of visual context depends also on the particularity of each specific concept; for instance, in Table V we observe that in the case of the semantic watershed algorithm and for the concepts *sea* and *person* the improvement measured over the complete dataset is 5% and 7.2%, respectively. Similarly, in the case of the semantic RSST and for concepts *sea, sand,* and *person* we see an overall increase significantly above the 2% average, namely 7.2%, 7.3%, and 14.6%, respectively.

Adding visual context to the segmentation algorithms is not an expensive process, in terms of computational complexity or timing. Average timing measurements for the contextualizing process on the set of 355 beach images illustrate that visual context is a rather fast process, resulting in an overall optimization of the results. Based on our implementation, initial color image segmentation resulting to approximate 30–40 regions requires about 10 s, while visual descriptors extraction and initial region labeling are the major bottleneck, requiring 60 and 30 s, respectively. Comparing to the above numbers, all proposed algorithms (semantic watershed, semantic RSST and visual context) have significantly lower computational time, in the order of one second. It is also worth noting, that context's resulting effect achieves an optimum of 13.10% increase, justifying its effectiveness in the semantic image segmentation process.

In order to test the robustness of our approach to imprecision in initial labels, we added several levels of Gaussian noise on the membership degrees of the initial region labeling of the images and repeated exactly the same experiment for semantic watershed segmentation to obtain a final evaluation score. We used a variety of values for the Gaussian noise variance, ranging from 0 (noise-free) to 0.30 with a step of 0.025. This variation scale was selected because it was observed that values above 0.30 produced erroneous results for all images, obtaining unfeasible membership degrees above 100%. The selection of the noise model is indifferent, since we want to test our algorithm with randomly altered input (i.e., partly incorrect initial labeling values) and this random alternation does not have to follow a specific distribution.

Application of Gaussian noise to the entire subset of 482 images resulted to the construction of the evaluation diagram presented in Fig. 9, illustrating the mean value for the evaluation score of each concept, as well as the overall evaluation score of the *beach* domain over different noise levels. As observed in Fig. 9 the overall behavior of our approach is stable and robust, considering minimal to medium amount of noise. More specifically, for small values of Gaussian noise, the total evaluation score remains at the same level as the noise-free score,

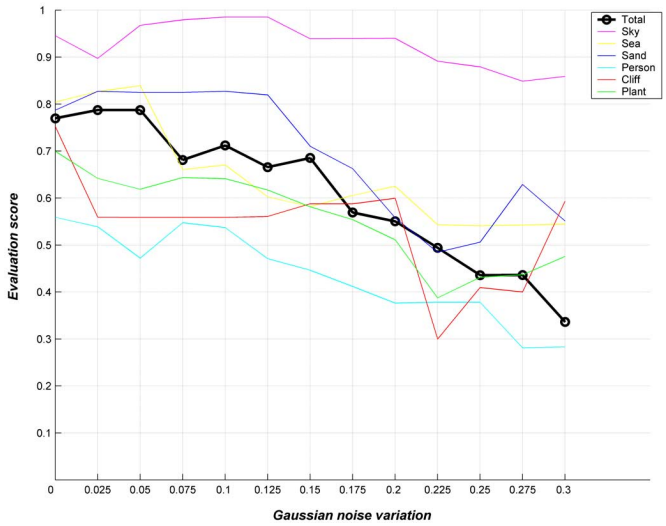

Fig. 9. Evaluation of semantic watershed's robustness against Gaussian noise over the entire data set of images.

approximately 70%–80%, while individual scores for each concept vary between 55%–95%, which is very good given the diversity of the image data set. Concepts *sky* and *sea* prove to have great resilience to noise, since we observe nearly stable and close to noise-free series of values even for great variance of Gaussian noise. In cases of very stressful noise, we expect that the proposed framework will conclude to nondeterministic behavior, which is verified by the evaluation score presented for high values of Gaussian noise addition; the evaluation score degrades and increases independently of the amount of noise added to the original images. However, provided that additional noise is kept to sane levels, the overall performance of the proposed contextual semantic region segmentation methodologies is decent.

VII. Conclusion

The methodologies presented in this paper can be exploited towards the development of more intelligent and efficient image analysis environments. Image segmentation and detection of materials and simple objects based on the semantic level, with the aid of contextual information, results into meaningful results. The core contributions of the overall approach have been the implementation of two novel semantic region growing algorithms, acting independently from each other, as well as a novel visual context interpretation based on an ontological representation, exploited towards optimization of the region label degrees of membership provided by the segmentation results. Another important point to consider is the provision of simultaneous still image region segmentation and labeling, providing a new aspect to traditional object detection techniques. In order to verify the efficiency of the proposed algorithms when faced with real-life data, we have implemented and tested them in the framework of developed research applications.

This approach made some interesting steps towards the correct direction and its developments are currently influencing subsequent research activities in the area of semantic-based image analysis. Future research efforts include tackling of composite objects in an image, utilizing both subgraphs and graphs instead of the straightforward approach of describing the image as a structured set of simple individual objects. Additionally, further exploitation of ontological knowledge is feasible by adding reasoning services as extensions to current approach. A fuzzy reasoning engine can compute fuzzy interpretations of regions, based on labels and fuzzy degrees, driving the segmentation process in a more structured way than for example the semantic distance of two neighbor regions used in this paper.

In this work, visual context aided to an extend to the semantic segmentation process (i.e., 7%–8% on average), however, it is the authors' belief that increased optimization can be achieved within a future, tweaked contextualization approach and our research efforts are focused on this field, as well. For instance, spatiotemporal relations may also be utilized during the contextualization step, whereas part of the proposed methodology may be used for detection of objects, i.e., incorporating alternative techniques in comparison to the graph matching technique currently utilized. Finally, conducting experiments over a much larger data set is within our priorities, as well as constructing proportionately ground truth information, while application of the proposed methodologies to video sequences remains always a challenging task.

References

[1] T. Adamek, N. O'Connor, and N. Murphy, "Region-based segmentation of images using syntactic visual features," presented at the Workshop Image Analysis for Multimedia Interactive Services (WIAMIS 2005) Montreux, Switzerland, Apr. 2005.

[2] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, Jun. 1994.

[3] G. Akrivas, G. Stamou, and S. Kollias, "Semantic association of multimedia document descriptions through fuzzy relational algebra and fuzzy reasoning," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 34, no. 2, pp. 190–196, Mar. 2004.

[4] G. Akrivas, M. Wallace, G. Andreou, G. Stamou, and S. Kollias, "Context—Sensitive semantic query expansion," in *Proc. IEEE Int. Conf. Artificial Intell. Syst. (ICAIS)*, Divnomorskoe, Russia, Sep. 2002, p. 109.

[5] T. Athanasiadis, V. Tzouvaras, K. Petridis, F. Precioso, Y. Avrithis, and Y. Kompatsiaris, "Using a multimedia ontology infrastructure for semantic annotation of multimedia content," presented at the 5th Int. Workshop Knowledge Markup and Semantic Annotation (SemAnnot'05), Galway, Ireland, Nov. 2005.

[6] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, *The Description Logic Hand-Book: Theory, Implementation and Application*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[7] A. Benitez, D. Zhong, S. Chang, and J. Smith, "MPEG-7 MDS content description tools and applications," in *Proc. ICAIP*, Warsaw, Poland, 2001, vol. 2124, p. 41.

[8] A. B. Benitez, "Object-based multimedia content description schemes and applications for MPEG-7," *Image Commun. J.*, vol. 16, pp. 235–269, 2000.

[9] S. Berretti, A. Del Bimbo, and E. Vicario, "Efficient matching and indexing of graph models in content-based retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 12, pp. 1089–1105, Dec. 2001.

[10] S. Beucher and F. Meyer, "The Morphological Approach to Segmentation: The Watershed Transformation," in *Mathematical Morphology in Image Processing*, E. R. Doughertty, Ed. New York: Marcel Dekker, 1993.

[11] M. Bober, "MPEG-7 Visual shape descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 716–719, Jun. 2001.

[12] E. Borenstein, E. Sharon, and S. Ullman, "Combining top-down and bottom-up segmentation," presented at the Proc. Comput. Vis. Pattern Recognit. Workshop, Washington, DC, Jun. 2004.

[13] M. Boutell and J. Luo, "Incorporating temporal context with content for classifying image collections," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR 2004)*, Cambridge, U.K., Aug. 2004, pp. 947–950.
[14] M. Boutell, J. Luo, X. Shena, and C. Brown, "Learning multilabel scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.
[15] S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent, "Unsupervised performance evaluation of image segmentation," *EURASIP J. Appl. Signal Process.*, vol. 2006, 2006, 96396.
[16] S.-F. Chang and H. Sundaram, "Structural and semantic analysis of video," in *IEEE Int. Conf. Multimedia Expo (II)*, 2000, p. 687.
[17] P. L. Correia and F. Pereira, "Objective evaluation of video segmentation quality," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 186–200, Feb. 2003.
[18] J. Dombi, "A general class of fuzzy operators, the De Morgan class of fuzzy operators and fuzziness measures induced by fuzzy operators," *Fuzzy Sets Syst.*, vol. 8, no. 2, pp. 149–163, 1982.
[19] H. Gao, W.-C. Siu, and C.-H. Hou, "Improved techniques for automatic image segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 12, pp. 1273–1280, Dec. 2001.
[20] T. R. Gruber, "A translation approach to portable ontology specification," *Knowledge Acquisition*, vol. 5, pp. 199–220, 1993.
[21] F. Jianping, D. K. Y. Yau, A. K. Elmagarmid, and W. G. Aref, "Automatic image segmentation by integrating color-edge extraction and seeded region growing," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1454–1466, Oct. 2001.
[22] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic, Theory and Applications*. Englewood Cliffs, NJ: Prentice Hall, 1995.
[23] B. Le Saux and G. Amato, "Image classifiers for scene analysis," presented at the Proc. Int. Conf. Comput. Vis. Graphics (ICCVG), Warsaw, Poland, Sep. 2004.
[24] J. Luo and A. Savakis, "Indoor versus outdoor classification of consumer photographs using low-level and semantic features," in *Proc. IEEE Int. Conf. Image Process.(ICIP01)*, 2001, vol. 2, pp. 745–748.
[25] A. Maedche, B. Motik, N. Silva, and R. Volz, "MAFRA—An Ontology Mapping Framework In The Context of the Semantic Web," presented at the Proc. Workshop Ontology Transform. ECAI2002, Lyon, France, Jul. 2002.
[26] N. Maillot, M. Thonnat, and A. Boucher, "Towards ontology based cognitive vision," *Mach. Vis. Appl.*, vol. 16, no. 1, pp. 33–40, Dec. 2004.
[27] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.
[28] C. Millet, I. Bloch, P. Hede, and P.-A. Moellic, "Using relative spatial relationships to improve individual region recognition," in *Proc. 2nd Eur. Workshop Integration Knowledge, Semantics and Digital Media Technol.*, 2005, pp. 119–126.
[29] S. Miyamoto, *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Norwell, MA: Kluwer, 1990.
[30] O. J. Morris, M. J. Lee, and A. G. Constantinides, "Graph theory for image analysis: An approach based on the shortest spanning tree," *Proc. Inst. Elect. Eng.*, vol. 133, pp. 146–152, Apr. 1986.
[31] *Information Technology—Multimedia Content Description Interface: Multimedia Description Schemes*, MPEG ISO/IEC FDIS 15938-5, ISO/IEC JTC1/SC29/WG11/M4242, Oct. 2001.
[32] P. Murphy, A. Torralba, and W. Freeman, "Using the forest to see the trees: A graphical model relating features, objects and scenes," in *Advances in Neural Information Processing Systems 16 (NIPS)*. Vancouver, BC: MIT Press, 2003.
[33] P. Mylonas and Y. Avrithis, "Context modeling for multimedia analysis and use," presented at the Proc. 5th Int. Interdiscipl. Conf. Modeling Using Context (CONTEXT'05), Paris, France, Jul. 2005.
[34] R. M. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 141–151, Mar. 2001.
[35] B. Neumann and R. Moeller, "On Scene Interpretation with Description Logics," in *Cognitive Vision Systems: Samping the Spectrum of Approaches*, H. I. Christensen and H.-H. Nagel, Eds. New York: Springer-Verlag, 2006, pp. 247–278.
[36] K. Petridis, S. Bloehdorn, C. Saathoff, N. Simou, S. Dasiopoulou, V. Tzouvaras, S. Handschuh, Y. Avrithis, I. Kompatsiaris, and S. Staab, "Knowledge representation and semantic annotation of multimedia content," *Proc. IEE Vis. Image Signal Processing, Special Issue on Knowledge-Based Digital Media Processing*, vol. 153, no. 3, pp. 255–262, Jun. 2006.
[37] P. Salembier and F. Marques, "Region-based representations of image and video—Segmentation tools for multimedia services," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1147–1169, Dec. 1999.
[38] L. Sanghoon and M. M. Crawford, "Unsupervised classification using spatial region growing segmentation and fuzzy training," in *Proc. IEEE Int. Conf. Image Process.*, Thessaloniki, Greece, 2001, pp. 770–773.
[39] J. P. Schober, T. Hermes, and O. Herzog, "Content-based image retrieval by ontology-based object recognition," in *Proc. KI-2004 Workshop Appl. Descript. Logics (ADL-2004)*, Ulm, Germany, Sep. 2004, pp. 61–67.
[40] S. Staab and R. Studer, *Handbook on Ontologies, International Handbooks on Information Systems*. Berlin, Germany: Springer-Verlag, 2004.
[41] *W3C, RDF*, [Online]. Available: http://www.w3.org/RDF/
[42] *W3C, RDF Reification*, [Online]. Available: http://www.w3.org/TR/rdf-schema/#ch_reificationvocab
[43] *W3C Recommendation, OWL Web Ontology Language Reference* , Feb. 10, 2004. [Online]. Available: http://www.w3.org/TR/owl-ref/

Thanos Athanasiadis (S'03) was born in Kavala, Greece, in 1980. He received the Diploma in electrical and computer engineering from the Department of Electrical Engineering, National Technical University of Athens (NTUA), Athens, Greece, in 2003, where he is currently working toward the Ph.D. degree at the Image Video and Multimedia Laboratory.

His research interests include knowledge-assisted multimedia analysis, image segmentation, multimedia content description, as well as content-based multimedia indexing and retrieval.

Phivos Mylonas (S'99) received the Diploma degree in electrical and computer engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 2001, the M.Sc. degree in advanced information systems from the National & Kapodestrian University of Athens (UoA), Athens, Greence, in 2003, and is currently pursuing the Ph.D. degree in computer science at NTUA.

He is currently a Researcher with the Image, Video and Multimedia Laboratory, School of Electrical and Computer Engineering, Department of Computer Science, NTUA, Greece. 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, utilizing fuzzy ontological knowledge aspects. He has published eight international journals and book chapters, is the author of 21 papers in international conferences and workshops, and has been involved in the organization of six international conferences and workshops.

Mr. Mylonas is a Reviewer for *Multimedia Tools and Applications*, *Journal of Visual Communication and Image Representation,* and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He is a member of the ACM, the Technical Chamber of Greece, and the Hellenic Association of Mechanical & Electrical Engineers.

Yannis Avrithis (M'95) was born in Athens, Greece, in 1970. He received the Diploma degree in electrical and computer engineering from the National Technical University of Athens (NTUA), Athens, Greence, in 1993, the M.Sc. degree in communications and signal processing (with distinction) from the Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, University of London, London, U.K., in 1994, and the Ph.D. degree in electrical and computer engineering from NTUA in 2001.

He is currently a Senior Researcher at the Image, Video and Multimedia Systems Laboratory, Electrical and Computer Engineering School, NTUA, conducting research in the area of semantic image and video analysis, coordinating research and development activities in national and European projects, and lecturing in NTUA. His research interests include spatiotemporal image/video segmentation and interpretation, knowledge-assisted multimedia analysis, content-based and semantic indexing and retrieval, video summarization, automatic and semi-automatic multimedia annotation, personalization, and multimedia databases. He has been involved in 13 European and 9 National R&D projects, and has published 23 articles in international journals, books, and standards, and 50 in conferences and workshops. He has contributed to the organization of 13 international conferences and workshops, and is a reviewer in 15 conferences and 13 scientific journals.

Dr. Avrithis is a member of ACM, EURASIP, and the Technical Chamber of Greece.

Stefanos Kollias (S'81–M'85) was born in Athens, Greece, in 1956. He received the Diploma in electrical and computer engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1979, the M.Sc. degree in communication engineering in 1980 from the University of Manchester Institute of Science and Technology, Manchester, U.K., and the Ph.D. degree in signal processing from the Computer Science Division, NTUA.

He has been with the Electrical Engineering Department, NTUA, since 1986, where he currently serves as a Professor. Since 1990, he has been the Director of the Image, Video, and Multimedia Systems Laboratory, NTUA. He has published more than 120 papers, 50 of which in international journals. He has been a member of the technical or advisory committee or invited speaker in 40 international conferences. He is a reviewer of ten IEEE Transactions and of ten other journals. Ten graduate students have completed their doctorate under his supervision, while another ten are currently performing their Ph.D. thesis. He and his team have been participating in 38 European and national projects.