A Context-based Region Labeling Approach for Semantic Image Segmentation[1]

Thanos Athanasiadis, Phivos Mylonas and Yannis Avrithis

School of Electrical and Computer Engineering
National Technical University of Athens
9, Iroon Polytechniou Str.,
157 73 Zographou, Athens, Greece
{thanos, fmylonas, iavr}@image.ntua.gr

Abstract. In this paper we present a framework for simultaneous image segmentation and region labeling leading to automatic image annotation. The proposed framework operates at semantic level using possible semantic labels 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 applied our idea on two region growing algorithms, i.e. watershed and recursive shortest spanning tree. Additionally we exploit the notion of visual context by employing fuzzy algebra and ontological taxonomic knowledge representation, incorporating in this way global information and improving region interpretation. In this process, semantic region growing labeling results are being re-adjusted appropriately, utilizing contextual knowledge in the form of domain-specific semantic concepts and relations. The performance of the overall methodology is demonstrated on a real-life still image dataset from the popular domains of beach holidays and motorsports.

1 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. Comparatively to the research efforts in automatic image and video segmentation [8], [18] and global [9], [14] or region-based [3], [13] image classification, still, human vision perception outperforms state-of-the-art computer algorithms. The main reason for this is that human vision is additionally based on high level a priori knowledge about the semantic meaning of the objects that compose the image and on contextual knowledge about their relationships. Moreover, erroneous image segmentation leads to poor results in recognition of materials and objects, while

[1] This research was partially supported by the European Commission under contract FP6-001765 aceMedia and contract FP6-027026 K-SPACE and by the Greek Secretariat of Research and Technology (PENED Ontomedia 03 EΔ 475).

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 work we propose an algorithm that involves simultaneously segmentation and detection of simple objects, imitating partly the way that human vision works. An initial region labeling is performed based on matching 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 concepts. 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, like morphological watershed [7], RSST [16], color-edge based and seeded region growing [11], etc., given some necessary modifications. Region growing algorithms start from an initial partition of the image and then an iteration of region merging begins, based on similarity measures until the predefined termination criteria are met. We adjust appropriately these merging process as well as the termination criteria.

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 utilized, combining fuzzy theory and fuzzy algebra [12] with characteristics derived from the Semantic Web, like the statement's reification technique [21]. In this process, membership degrees of concepts assigned to regions derived by the semantic segmentation process are optimized, according to a context-based membership degree readjustment algorithm. This algorithm utilizes ontological knowledge, in order to provide optimized membership degrees of detected concepts of each region in the scene. Our research efforts employ contextual knowledge derived from the popular domains of beach holidays and motorsports.

The outline of the paper is as follows: Section 2 is dedicated to the knowledge representation used, including the necessary notation used throughout the paper. Section 3 describes the semantic region growing approach of segmentation, examining in detail two variations. Utilization of contextual knowledge is discussed in section 4 and finally section 5 presents the dataset and methodology of the experiments and the results of the proposed algorithms.

2 Knowledge Representation

2.1 Ontology fuzzification and fuzzy relations

The first thing to consider within the proposed approach of semantic image segmentation and labeling is what type of knowledge model to use to describe the contextual information. The latter plays a key role in optimizing the results of both methodologies and is built on a novel ontological representation for context. In general, one possible way to describe ontologies [10] can be formalized as:

$$O = \{C, \{R_{c_i,c_j}\}\} \tag{1}$$

O is an ontology, C is the set of concepts described by the ontology, c_i and c_j are two concepts $c_i, c_j \in C$ and $R_{c_i,c_j}: C \times C \to \{0,1\}$ is the semantic relation amongst these concepts, as the latter is defined within the semantic framework of the MPEG-7 description [20]. According to this description narrative worlds depicted by or related to multimedia content are represented by describing semantic concepts together with their relations and attributes [6]. Herein, the proposed knowledge model is based on a set of concepts and relations between them, which form the basic elements towards semantic interpretation. Although almost any type of relation may be included to construct the knowledge representation, the two main categories used are *taxonomic* (i.e. ordering) and *compatibility* (i.e. symmetric) relations. However, compatibility relations fail to assist in the determination of the context and therefore the use of ordering relations is more appropriate for such tasks [1]. Thus, a 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. However, in this case available information will be scattered among them, making each one of them inadequate to describe a context in a meaningful way. Thus, the utilized 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 MPEG-7 [19], namely the *specialization* relation Sp, the *part* relation P and the *property* relation Pr. When modeling real-life information that is governed by uncertainty and fuzziness, fuzzy relations have been proposed to handle such issues. In particular, the above commonly encountered 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 types of relations we propose a "fuzzification" of the previous ontology definition, as follows:

$$O_F = \{C, \{r_{c_i,c_j}\}\}, \text{ where } r_{c_i,c_j} = F(R_{c_i,c_j}) : C \times C \to [0,1] \tag{2}$$

In equation (2), O_F defines a "fuzzified" ontology, C is again the set of all possible concepts it describes and r_{c_i,c_j} denotes a fuzzy relation amongst the two concepts $c_i, c_j \in C$. More specifically, given a universe U a crisp set C is described by a membership function $\mu_C : U \to \{0,1\}$, whereas according to [12], a *fuzzy set* F on C is described by a membership function $\mu_F : C \to [0,1]$. We may describe the fuzzy set F using the sum notation:

$$F = \sum_{i=1}^{n} c_i / w_i = \{c_1 / w_1, c_2 / w_2, \ldots, c_n / w_n\} \tag{3}$$

Where $n = |C|$ is the cardinality of C and $w_i = \mu_F(c_i)$. As in [12], a *fuzzy relation* on C is a function $r_{c_i,c_j} : C \times C \to [0,1]$ and its *inverse* relation is defined as $r_{c_i,c_j}^{-1} = r_{c_j,c_i}$. Based on the relations r_{c_i,c_j} 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*.

$$T = Tr^t(Sp \bigcup P^{-1} \bigcup Pr^{-1}) \tag{4}$$

Transitive closure Tr^t is required in order for T to be taxonomic, as the union of transitive relations is not necessarily transitive [2].

2.2 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) 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, non-overlapping 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 $v_a \equiv \langle a, D_a, L_a \rangle$. More specifically, $D_a = [DC_a \ \mathrm{HT_a}]$ is the ordered set of two MPEG-7 Visual Descriptors characterizing the region in terms of low-level features, namely *Dominant Color* (*DC*) and *Homogeneous Texture* (*HT*) [15]. Additionally, $L_a = \sum_{i=1}^{|C|} c_i / \mu_a(c_i)$ is the fuzzy set (defined on the crisp set of concepts C, since $c_i \in C$) of candidate concepts for the region, which incorporates the uncertainty of the of the region labeling process.

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 pair of adjacent regions (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:

$$s_{ab} = \max_{c_k \in C}(\min(\mu_a(c_k), \mu_b(c_k))),\ a,b \in G \tag{5}$$

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 $\mu_a(c_k)$ and $\mu_b(c_k)$ 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 ARG, a neighborhood N_a of a vertex $v_a \in V$ is the set of vertices whose corresponding regions are connected to *a*: $N_a = \{v_b : e_{ab} \neq \varnothing\},\ a,b \in G$. 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.

3 Semantic Region Growing Approach

3.1 Overview

The major target of this work is to improve both image segmentation and recognition of 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 mid-level features, in the formal style defined earlier in section 2.2. Our intention is to operate on a higher level of information where regions are linked to concepts rather than only to their visual features. For this purpose a knowledge assisted analysis (KAA) algorithm, discussed in depth in a previous work [4], has been designed and implemented. Population of the fuzzy set L_a for all regions of *G*, is based on a matching process between the visual descriptors stored in each vertex v_a of the ARG and the corresponding visual descriptors of concepts, stored in the form of prototype instances in the corresponding ontological knowledge base.

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 [7], while the second is the Recursive Shortest Spanning tree, also known as RSST [16]. 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.*) [5]. 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).

3.2 Semantic Watershed

The watershed algorithm [7] owes its name from 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. A subset of regions $S \subseteq G$ is selected to be used as seeds for the initialization of the semantic watershed algorithm. The criteria for selecting a region to become a seed, i.e. $s \in S$, are the following two:

1. The height of its fuzzy set L_s (maximum degree of membership in the fuzzy set [12]) should be above a threshold: $h(L_s) \equiv \max_{c_k \in C}(\mu_s(c_k)) > T_{seed}$. Threshold T_{seed} is calculated once in the beginning of the algorithm, based on the histogram of all degrees of membership over all regions of the image.
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:

$$h(L_s) > \sum_{c_k \in \{C - c^*\}} \mu_s(c_k), \text{ where } c^* : \mu_s(c^*) = h(L_s) \tag{6}$$

These two constrains ensure that the specific region has been correctly selected as seed for the particular concept c^*.

An iterative process begins checking for every initial region-seed $s \in S$ in all its direct neighbors $n \in N_s$ (as defined in the ARG) if they have been assigned to the same concept c and, with what degree of membership $\mu_n(c_k)$. Some of those regions, that satisfy an additional criterion, form a new set of regions M^i (i denotes the iteration step, with $M^0 \equiv S$), which will be the new seeds for the next iteration of the algorithm. These additional criterion is that the degree of membership of region n under examination, for the particular concept c should be above a merging threshold: $\mu_n(c_k) > K^i \cdot T_{merge}$, where K is a constant slightly above one, that increases the threshold in every iteration i of the algorithm in a non linear way to the distance from the initial regions-seeds. When the above criteria are satisfied, region n is merged with its propagator s and an updated degree of membership is calculated using the default t-norm for the newly created region:

$$\mu_{\hat{s}}(c_k) = \min(\mu_s(c_k), \mu_n(c_k)) \tag{7}$$

The termination criterion of the algorithm is quite straightforward: repeat this procedure until the set of regions-seeds in step i is empty: $M^i = \varnothing$. In 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

3.3 Semantic RSST

Traditional RSST [16] 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 2.2. 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. Then recursively 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 re-calculated. Traditional RSST will also re-calculate weights of related edges as well and resort them, so that in every step the edge with the least weight will be selected.

Following the conventions and notation used so far, we introduce here a modified version of RSST, called Semantic RSST (S-RSST). The first step is to populate the set of edges *E* by traversing the ARG. 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 similaruty value defined earlier in section 2.2. Commutativity and associativity axioms of all fuzzy set operations (thus including default t-norm and default s-norm) ensure that the ordering of the arguments is indifferent. In this way all graph's edges are sorted by their weight:

$$w(e_{ab}) = 1 - s_{ab} \tag{8}$$

Equation (8) can be expanded by substituting s_{ab} from equation (5). 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.

Let us now examine in details one iteration of the S-RSST algorithm. Firstly, the edge with the least weight is selected as:

$$e^*_{ab} = \arg\min_{e_{ab} \in E}(w(e_{ab})) \tag{9}$$

Then regions *a* and *b* are merged to form a new region $\hat{a}$. Vertex v_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:

$$\mu_{\tilde{a}}(c_k) = \frac{A(a)\cdot\mu_a(c_k) + A(b)\cdot\mu_b(c_k)}{A(a)+A(b)},\ \forall c_k \in C \tag{10}$$

The quantity $A(a)$ is a measure of the size of *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*: $e \in E_a \bigcup E_b$.

This procedure continues until the edge e^* with the least weight in the ARG is bigger than a threshold: $w(e^*) > T_w$. This threshold is calculated in the beginning of the algorithm, based on the cumulative histogram of the weights of *E*.

4 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 [17], we restrict it herein to the notion of ontological context. The latter is defined within the "fuzzified" version of traditional ontologies presented in section 2.1 and 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 its best performance.

Based on the mathematical foundations described in previous subsections, we introduce the algorithm used to readjust the degree of membership $\mu_a(c_k)$ of each concept $c_k \in C$ associated to a region $a \in G$ of the scene. Each specific concept c_k is present in the application-domain's ontology, stored together with its relationship degrees r_{c_k,c_j} to any other related concept c_j. To tackle cases that more than one concept is related to multiple concepts, the term *context relevance* $cr_{dm}(c_k)$ is introduced, which refers to the overall relevance of concept c_k to the *root element* characterizing each domain dm. For instance the root element of beach and motorsports domains are concepts c_{beach} and $c_{motorsport}$ 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 one that quantifies their semantic correla-

tion. For the problem at hand and given both the beach and motorsports domains, the *max* value is a meaningful measure of correlation. A simplified example, limiting the only available concepts to $c_{motorsport} = c_m$, $c_{asphalt} = c_a$, $c_{grass} = c_g$ and $c_{car} = c_c$ is presented in Fig. 1 and summarized in the following: letting concept c_a be related to concepts c_m, c_g and c_c directly with: r_{c_a,c_m}, r_{c_a,c_g} and r_{c_a,c_c}, while concept c_g is related to concept c_m with r_{c_g,c_m} and concept c_c is related to concept c_m with r_{c_c,c_m}. Additionally, c_c is related to c_g with r_{c_c,c_g}. Then, we calculate the value for $cr_{dm}(c_a)$:

$$cr_{dm}(c_a) = \max \begin{Bmatrix} r_{c_a,c_m}, r_{c_a,c_g} \cdot r_{c_g,c_m}, r_{c_a,c_c} \cdot r_{c_c,c_m}, \\ r_{c_a,c_g} \cdot r_{c_g,c_c} \cdot r_{c_c,c_m}, r_{c_a,c_c} \cdot r_{c_c,c_g} \cdot r_{c_g,c_m} \end{Bmatrix} \quad (11)$$

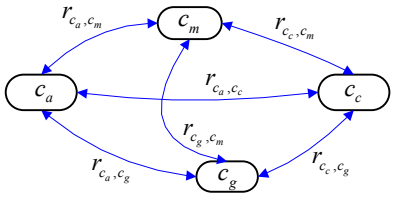

Fig. 1. Graph representation example – Compatibility indicator estimation

The general structure of the proposed re-evaluation algorithm is summarized in the following steps:

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 $np \to [0,1]$.
2. For each concept $c_k \in C$ in the fuzzy set L_a associated to a region $a \in G$ in a scene with a degree of membership $\mu_a(c_k)$, obtain the particular contextual information in the form of its relations to the set of any other concepts: $\{r_{c_k,c_j} : c_j \in C, c_j \neq c_k\}$.
3. Calculate the new degree of membership $\mu_a(c_k)$ 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 c_k to more than one concepts, rather than relating c_k solely to the "root element" c_r, an intermediate aggregation step should be applied for c_k: $cr_{c_k} = \max\{r_{c_k,c_r},..,r_{c_k,c_m}\}$. We express the calculation of $\mu_a(c_k)$ with the recursive formula:

$$\mu_a^i(c_k) = \mu_a^{i-1}(c_k) - np \cdot (\mu_a^{i-1}(c_k) - cr_{c_k}) \quad (12)$$

Where *i* denotes the iteration used. Equivalently, for an arbitrary iteration *i*:

$$\mu_a^i(c_k) = (1-np)^t \cdot \mu_a^0(c_k) + (1-(1-np)^i) cr_{c_k} \quad (13)$$

Where $\mu_a^0(c_k)$ represents the original degree of membership. Typical values for *i* reside between 3 and 5.

Key point in this approach remains the definition of a meaningful normalization parameter np. When re-evaluating this value, the ideal np is always defined with respect to the particular domain of knowledge and is the one that quantifies each semantic correlation to the domain. In this work we conducted a series of experiments on a "training" subset of 52 images for both application domains and selected the np that resulted in the best overall precision/recall values for each domain.

5 Experimental Results

We carried out experiments in the domains of beach and motorsports, utilizing 262 images in total, i.e. 193 beach and 69 motorsports images acquired either from the internet or from personal collections. In order to demonstrate the proposed methodologies and keep track of each individual algorithm results, we integrated the described techniques into a single application that utilizes a user-friendly graphical interface. In the following we present two representative sets of experimental results, i.e. one image derived from the beach domain and one image from the motorsports domain. Each set includes four images: (a) the original image, (b) the result of traditional RSST, (c) the result of semantic watershed and (d) the result of semantic RSST. In the case of the traditional RSST, we pre-defined the final number of regions to be produced to be equal to the ones produced by the semantic watershed; in this fashion segmentation results are easily comparable.

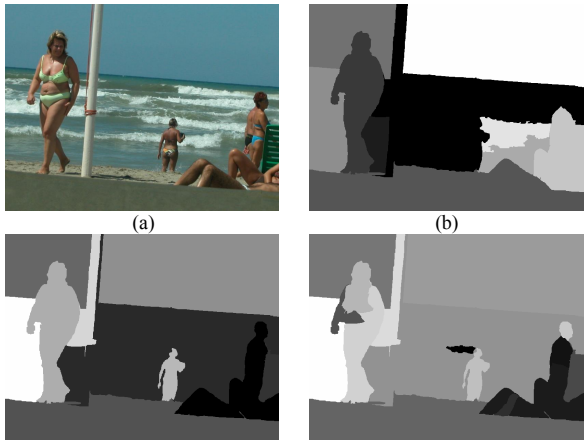

Fig. 2. Experimental results for the beach domain – Example 2. (a) Input image, (b) RSST segmentation, (c) semantic watershed, (d) semantic RSST

Fig. 2 illustrates the example derived from the beach domain. As observed in Fig. 2b, RSST segmentation results are insufficient: some persons are merged 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. 2c). Sea regions on the left part of the image are successfully merged together, the woman on the left is correctly identified as one region, successfully tackling the existence of variations in low level characteristics, i.e. green swimsuit vs. 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, semantic RSST algorithm in Fig. 2d performs similarly well. Small differences with semantic watershed are justified by the fact that in S-RSST 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.

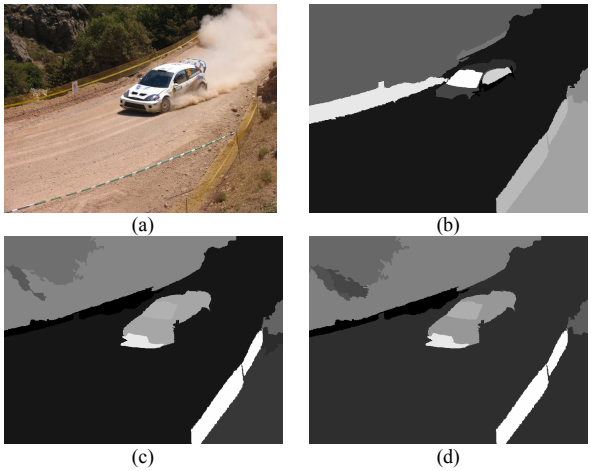

Fig. 3. Experimental results for the motorsports domain. (a) Input image, (b) RSST segmentation, (c) semantic watershed, (d) semantic RSST

Results from the motorsports domain are described in Fig. 3. More specifically, in Fig. 3a we present the original image derived from the World Rally Championship. Plain segmentation results (Fig. 3b) 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. 3c and Fig. 3d 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 SW and SRSST approaches is summarized in the way they handle vegetation in the upper left corner of the image, with SRSST performing closer to the ground truth, since it detects the variations in vegetation and grass successfully.

Finally, we continue with presenting a 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 confidence values.

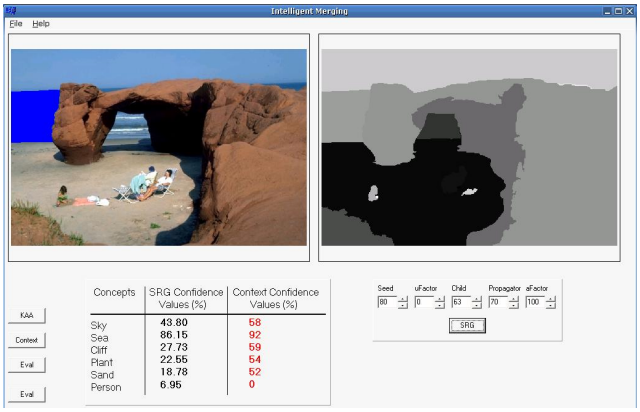

Fig. 4. Contextual experimental results for the first beach image example

In Fig. 4 we observe the contextualization step for the first beach image, presented in the Contextual Analysis tool developed. Contextualization, which works on a per region basis, is applied after the semantic region merging, 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 blue color. The contextualized results are presented in red in the right text column at the bottom of the GUI. Context favors strongly the fact that the merged region belongs to sea, increasing its confidence value from 86.15% to a crisp 92%. Additionally, the totally irrelevant (for the specific region) confidence value for the concept person is extinguished, whereas medium confidence values for the rest of the possible beach concepts are slightly increased, due to the ontological knowledge relations that exist in the considered knowledge model. That is because of the relationships that exist in the a priori built contextual knowledge and that strongly relate concepts encountered on a beach scene with each other, we expect that the use

of context will improve the results but at the same time provide also some false concepts as well. However, in all cases context does normalize results in a meaningful manner, i.e. each region's dominant concept is detected in comparison to ground truth and its degree of membership is increased.

6 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 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 region' associated fuzzy set of concepts 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.

References

[1] G. Akrivas, G. Stamou and S. Kollias, "Semantic Association of Multimedia Document Descriptions through Fuzzy Relational Algebra and Fuzzy Reasoning", IEEE Trans. on Systems, Man, and Cybernetics, part A, Volume 34 (2), March 2004

[2] G. Akrivas, M. Wallace, G. Andreou, G. Stamou and S. Kollias, "Context – Sensitive Semantic Query Expansion", Proc. of the IEEE International Conference on Artificial Intelligence Systems (ICAIS), Divnomorskoe, Russia, September 2002

[3] E. L. Andrade Neto, J. C. Woods, E. Khan, M.Ghanbari "Region Based Analysis and Retrieval for Tracking of Semantic Objects and Provision of Augmented Information in Interactive Sport Scenes" IEEE Trans. on Multimedia Vol. 7, Issue 6, pp. 1084–1096, December 2005

[4] Th. Athanasiadis, V. Tzouvaras, K. Petridis, F. Precioso, Y. Avrithis and Y. Kompatsiaris, "Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content", Proc. of 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot '05), Galway, Ireland, November 2005

[5] Th. Athanasiadis, Y. Avrithis, S. Kollias, "A Semantic Region Growing Approach in Image Segmentation and Annotation", 1st International Workshop on Semantic Web Annotations for Multimedia (SWAMM), Edinburgh, Scotland, November 2006

[6] A. B. Benitez, H. Rising, C. Jrgensen, R. Leonardi, A. Bugatti, K. Hasida, R. Mehrotra, M. Tekalp, A. Ekin, and T. Walker, "Semantics of Multimedia in MPEG-7", In IEEE International Conference on Image Processing, pages 137--140, vol.1, 2002

[7] S. Beucher and F. Meyer, "The Morphological Approach to Segmentation: The Watershed Transformation", Mathematical Morphology in Image Processing, E.R.Doughertty (Ed.), Marcel Dekker, NY, 1993

[8] E. Borenstein, E. Sharon and S. Ullman, "Combining Top-Down and Bottom-Up Segmentation", Computer Vision and Pattern Recognition Workshop, Washington DC, USA, June 2004

[9] M. Boutell, J. Luo, X. Shena and C. Brown, "Learning multi-label scene classification", Pattern Recognition, 37(9), pp. 1757-1771, September 2004
[10] T.R. Gruber, "A Translation Approach to Portable Ontology Specification", Knowledge Acquisition 5: 199-220, 1993
[11] 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. on Image Processing, Vol. 10, No. 10, pp. 1454-1466, October 2001
[12] G. Klir and B. Yuan, "Fuzzy Sets and Fuzzy Logic, Theory and Applications", New Jersey, Prentice Hall, 1995
[13] S. Lee, M. M. Crawford, "Unsupervised classification using spatial region growing segmentation and fuzzy training", Proc. of the IEEE International Conference on Image Processing, Thessaloniki, Greece, 2001
[14] J. Luo and A. Savakis, "Indoor vs outdoor classification of consumer photographs using low-level and semantic features", In Proc. IEEE Int. Conf. on Image Processing (ICIP01), 2001
[15] B.S. Manjunath, J.R. Ohm, V.V. Vasudevan, A. Yamada, "Color and texture descriptors", Special Issue on MPEG-7, IEEE Trans. on Circuits and Systems for Video Technology, 11/6, 703-715, June 2001
[16] 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, April 1986, pp. 146–152
[17] Ph. Mylonas and Y. Avrithis, "Context modeling for multimedia analysis and use", Proc. of 5th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT '05), Paris, France, July 2005
[18] P. Salembier, F. Marques, "Region-Based Representations of Image and Video - Segmentation Tools for Multimedia Services", IEEE Trans. on Circuits and Systems for Video Technology, vol.9, no.8, 1999
[19] T. Sikora, "The MPEG-7 Visual standard for content description - an overview", Special Issue on MPEG-7, IEEE Trans. on Circuits and Systems for Video Technology, 11/6:696-702, June 2001
[20] ISO/IEC JTC 1/SC 29/WG 11/N3966, "Text of 15938-5 FCD Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes", Singapore, 2001
[21] W3C, RDF Reification, http://www.w3.org/TR/rdf-schema/#ch_reificationvocab