Knowledge Driven Segmentation and Classification

Thanos Athanasiadis[1], Phivos Mylonas[1], Georgios Th. Papadopoulos[2], Vasileios Mezaris[2], Yannis Avrithis[1], Ioannis Kompatsiaris[2], and Michael G. Strintzis[2]

[1] Image, Video and Multimedia Systems Laboratory, National Technical University of Athens, 15780 Zographou, Greece,
(thanos,fmylonas,iavr)@image.ntua.gr
[2] Informatics and Telematics Institute / Centre for Research and Technology Hellas, 57001 Thermi-Thessaloniki, Greece,
(papad,bmezaris,ikom,michael)@iti.gr

1 Introduction

In this chapter a first attempt will be made to examine how the coupling of multimedia processing and knowledge representation techniques, presented separately in previous chapters, can improve analysis. No formal reasoning techniques will be introduced at this stage; our exploration of how multimedia analysis and knowledge can be combined will start by revisiting the image and video segmentation problem. Semantic segmentation, presented in the first section of this chapter, starts with an elementary segmentation and region classification and refines it using similarity measures and merging criteria defined at the semantic level. Our discussion will continue in the next sections of the chapter with knowledge-driven classification approaches, which exploit knowledge in the form of contextual information for refining elementary classification results obtained via machine learning. Two relevant approaches will be presented. The first one deals with visual context and treats it as interaction between global classification and local region labels. The second one deals with spatial context and formulates the exploitation of it as a global optimization problem. All approaches presented in this chapter are geared towards the "Photo Use Case" scenario defined in Chapter 2, although their use as part of a semi-automatic annotation process in a Professional Media Production and Archiving setting is also feasible.

2 Related Work

Starting from an initial image segmentation and the classification of each resulting segment in one of a number of possible semantic categories, using one of the various segmentation algorithms of the literature and one or more of the classifiers discussed in Chapter 5, there are two broad categories of possible analysis errors that one may encounter: segmentation errors and classifications errors. Segmentation errors occur as either under-segmentation or over-segmentation; in both cases, the result is the formation of one or more spatial regions, each of which does not accurately correspond to a single semantic object depicted in the image. Classification errors, on the other hand, occur as a result of the insufficiency of the employed combination of classification technique and feature vector to effectively distinguish between different classes of objects. Clearly, these two categories of analysis errors are not independent of each other: a segmentation error such as the formation of regions that correspond to only a small part of a semantic object, for example, can clearly render useless any employed shape features and thus lead to erroneous classification; there are several similar examples of how segmentation affects classification performance.

In order to minimize the number of segmentation errors, several elaborate image segmentation methods have appeared in the relevant literature. Some of them focus on the use of a more complete set of visual cues for performing segmentation, e.g. the combined use of color, texture and position features [14], and the introduction of new algorithms for exploiting these features, while others attempt to minimize segmentation errors by means of post-processing procedures that perform region-merging or even region-splitting operations upon an appropriately generated initial segmentation [1]. The introduction of semantics-based criteria in the approaches of the latter category is an interesting idea that can contribute towards better segmentation; a method for this is presented in the sequel. A comprehensive review of image segmentation methods not exploiting semantic information is nevertheless beyond the scope of this chapter; the interested reader is referred to [15] for a review on this topic.

To allow for the more reliable classification of the generated regions, on the other hand, and in particular to address the insufficiency of a given combination of classification technique and feature vector to effectively distinguish between different classes of objects, the use of contextual information has been proposed. Contextual information, for the purpose of semantic image analysis, can refer to spatial information, concept co-occurrence information and other kinds of prior knowledge that can contribute to the disambiguation of the semantics of a spatial region based on the semantics of its peers [12].

Spatial information in particular has been shown to be very suitable for discriminating between objects exhibiting similar visual characteristics, since it is generally observed that objects tend to be present in a scene within a particular spatial context [21]. To this end, several approaches have been proposed

in the relevant literature that utilize spatial information in order to overcome the ambiguities and limitations that are inherent in the visual medium. In [9], Hollink et al. discusses the issue of semi-automatically adding spatial information to image annotations. Among the most commonly adopted spatial context representations, directional spatial relations have received particular interest. They are used to denote the relative position of objects in space and their capability in facilitating semantic image analysis tasks has been highlighted. The relevant literature considers roughly of two categories for the latter: angle-based and projection-based approaches. Angle-based approaches include [26], where a pair of fuzzy k-NN classifiers are trained to differentiate between the *Above/Below* and *Left/Right* relations and the work of [16], where an individual fuzzy membership function is defined for every relation and applied directly to the estimated angle-histogram. Projection-based approaches include [9], where qualitative directional relations in terms of the center and the sides of the corresponding objects' Minimum Bounding Rectangles (MBRs) were defined.

3 Semantic Image Segmentation

3.1 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*) [6] 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 $v_a \equiv \langle a, D_a, L_a\rangle$. 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 $e_{ab} = \langle (v_a, v_b), s_{ab}\rangle \in E$. 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:

$$s_{ab} = \sup_{c \in C}(t_{norm}(\mu_a(c), \mu_b(c))), a, b \in G \quad (1)$$

The above formula states that the similarity of two regions is the supremum (sup) over all common concepts of the fuzzy intersection (t_{norm}) of the

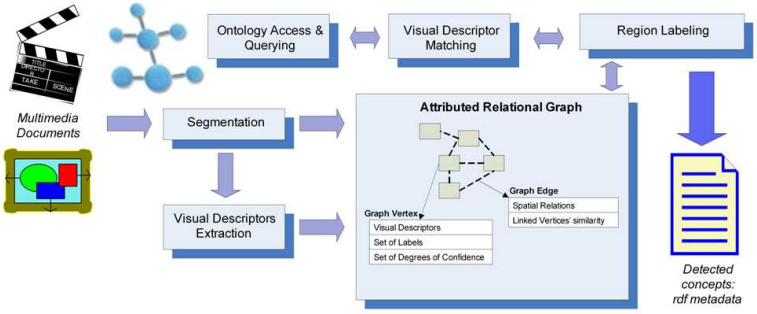


Fig. 1. Initial region labeling based on ARG and Visual Descriptors matching.

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: $E_a = \{e_{ab} : b \in N_a\} \subseteq E$.

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.

3.2 Image Graph Initialization

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) [13], 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 [4], has been designed and implemented. The general architecture scheme is depicted in Fig. 1, where in the center lies the ARG, interacting with the rest of the processes.

The ARG is constructed based on an initial RSST-like segmentation [1] that produces a few tens of regions (approximately 30-40 in our experiments). For every region *Dominant Color* (DC) and *Homogeneous Texture* (HT) are extracted (i.e. for region a: $D_a = [DC_a HT_a]$) and stored in the corresponding graph's vertex. The formal definition of DC [13] is $DC \equiv [\{c_i, v_i, p_i\}, s], i = 1..N$, where c_i is the i$^{\text{th}}$ 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 DC_1, DC_2 is:

$$d_{DC}(DC_1, DC_2) = \sqrt{\sum_{i=1}^{N_1} p_{1i}^2 + \sum_{j=1}^{N_2} p_{2j}^2 - \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} 2a_{1i,2j}p_{1i}p_{2j}} \quad (2)$$

where $a_{1i,2j}$ is a similarity coefficient between two colors. Similarly for HT we have $HT \equiv [avg, std, e_1, .., e_{30}, d_1, .., d_{30}]$, where avg is the average intensity of the region, std is the standard deviation of the region's intensity, 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:

$$d_{HT}(HT_1, HT_2) = \sum_{i=1}^{N_{HT}=62} \left| \frac{HT_1(i) - HT_2(i)}{\sigma_i} \right| \quad (3)$$

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: $d_{DC}(DC_a, DC_b) \equiv d_{DC}(a, b)$ (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:

$$\begin{aligned} d(a, p) &= d([DC_a HT_a], [DC_p HT_p]) \\ &= w_{DC}(c) \cdot n_{DC}(d_{DC}(a, p)) + w_{HT}(c) \cdot n_{HT}(d_{HT}(a, p)) \end{aligned} \quad (4)$$

where d_{DC} and d_{HT} are given in equations (2) and (3), w_{DC} and w_{HT} are weights depending on each concept c and $w_{DC}(c) + w_{HT}(c) = 1, \forall c \in C$. Additionally, n_{DC} and n_{HT} are normalization functions and more specifically were selected to be linear:

$$n(x) = \frac{x - d_{\min}}{d_{\max} - d_{\min}}, n : [d_{\min} d_{\max}] \rightarrow [01] \quad (5)$$

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:

$$\mu_a(c) = 1 - \min_{p \in P(c)} d(a, p) \quad (6)$$

where $d(a, p)$ is given in (4). 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 chapter suggesting an extension based on region merging and segmentation on a semantic level.

3.3 Semantic Region Growing

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 mid-level features, like the fuzzy sets of labels we defined earlier in section 3.1.

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 [19]. 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).

Semantic Watershed

The watershed algorithm [7] 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 the following two:

1. The height of its fuzzy set L_a (the largest degree of membership obtained by any element of L_a [11]) should be above a threshold: $h(L_a) > T_{seed}$. 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 derived from a training set of images.
2. The specific region has only one dominant concept, i.e. the remaining concepts should have low degrees of membership comparatively to that of the dominant concept:

$$h(L_a) > \sum_{c \in \{C - c*\}} \mu_a(c), where c* : \mu_a(c*) = h(L_a) \tag{7}$$

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 $\mu_r(c) > K^n \cdot T_{merge}$, 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 non-linear 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 $M_s^0 \equiv s$, $M_s^1 \equiv N_s$, 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:

$$\mu_{\hat{r}}(c) = \min(\mu_{p(r)}(c), \mu_r(c)) \tag{8}$$

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 repeatedly run our algorithm on new *ARGs*; each one of the latter consists of the specific 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.

Semantic RSST

Traditional RSST [19] 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 3.1. 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 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. 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 3.1. For an edge e_{ab} between two adjacent regions a and b we define its weight as follows:

$$w(e_{ab}) = 1 - s_{ab} \tag{9}$$

Equation (9) can be expanded by substituting s_{ab} from equation (1). 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: $e^*_{ab} = \arg\min_{e_{ab} \in E} (w(e_{ab}))$.

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

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

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:

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

   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: $e \in E_a \bigcup E_b$.

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.

Fig. 2 illustrates an example derived from the beach domain. In order to make segmentation results comparable we pre-defined the final number of regions produced by traditional RSST to be equal to that produced by the semantic watershed. An obvious observation is that RSST segmentation performance in Fig. 2b is rather poor; persons are merged with sand, whereas sea on the left and in the middle under the big cliff is divided into several regions and adjacent regions of the same cliff are classified as different ones. The results of the application of the semantic watershed algorithm are shown in Fig. 2c and are considerably better. More specifically, we observe that both parts of sea are merged together, and the rocks on the right side are represented by only two large regions, despite their original variations in texture and color information. Moreover, the persons lying on the sand are identified as separate regions. Semantic RSST (Fig. 2d) is shown to perform similarly well.

4 Using Contextual Knowledge to Aid Visual Analysis

4.1 Contextual Knowledge Formulation

Ontologies [22] present a number of advantages over other knowledge representation strategies. 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 [5] or OWL [25]). One possible way to describe ontologies can be formalized as:

$$O = \{C, \{R_{pq}\}\}, \text{where } R_{pq} : C \times C \rightarrow \{0, 1\} \tag{11}$$

where 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 above knowledge model encapsulates a set of concepts and the relations between them, forming the basic elements toward semantic interpretation. In general, semantic relations describe specific kinds of links or relationships between any two concepts. In the crisp (non-fuzzy) 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 [2], compatibility relations fail to assist in the determination of the context and the use of ordering relations is necessary for such tasks. Thus, the first challenge is to meaningfully use information from the taxonomic relations to exploit context for semantic image segmentation and object labeling.

For a knowledge model to be highly descriptive, it must contain many distinct and diverse relations among its concepts. A side-effect of using many diverse relations is that available information will then be scattered across them, making any individual relation inadequate for describing 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 use three types of relations, whose semantics are defined in the MPEG-7 standard [10], namely the *specialization* relation Sp, the *part of* relation P and the *property* relation Pr.

One more important point must be considered when designing a knowledge model: real-life data is often considerably different from research data. Real-life information is, in principal, governed by uncertainty and fuzziness. It can therefore be more accurately modelled using fuzzy relations. The commonly encountered crisp relations above can be modeled as fuzzy ordering relations, and can be combined to generate a meaningful fuzzy taxonomic relation. To

tackle such complex types of relations we propose a "fuzzification" of the previous ontology definition:

$$O_F = \{C, \{r_{pq}\}\}, \text{where } r_{pq} = F(R_{pq}) : C \times C \rightarrow [0,1] \tag{12}$$

where O_F defines a fuzzy ontology, C is again the set of all possible concepts it describes and r_{pq} denotes a fuzzy relation amongst the 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 [11], 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 [18]:

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

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 [11], 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 purposes of image analysis here, we construct a relation T using the transitive closure of the fuzzy taxonomic relations: *Specialization* Sp, *Part of* P and *Property* Pr:

$$T = Tr^t(Sp \cup P^{-1} \cup Pr^{-1}) \tag{13}$$

In these 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 likely 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 [10] contains several types of semantic relations meaningful to multimedia analysis, 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. There is, similarly, a meaningful inverse for the *property* relation Pr. 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 mammal and q 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 (13) 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. 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 [3].

Representation of our concept-centric contextual knowledge model follows the Resource Description Framework (RDF) standard [23]. 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 that express a relationship between the subject and the object. The relation T can be visualized as a graph, in which every node represents a concept and each edge constitutes a contextual relation between these concepts. Additionally each edge has an associated membership degree, which represents the fuzziness within the context model. Representing the graph in RDF is straightforward, since RDF structure itself is based on a similar graph model.

To represent fuzzy relations, we use reification [24]: a method for making statements about other statements in RDF. In our model, the reified statements capture the degree of membership for the relations. This method of representing fuzziness 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 not imply that sky is always a part of a beach scene.

A small clarifying example is provided in Fig. 3 for an instance of the specialization relation Sp. As discussed, $Sp(x, y) > 0$ implies that the meaning of x "includes" the meaning of y; the most common forms of specialization are sub-classing, 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. The reification process introduces a statement about the *specializationOf* predicate, stating that the membership degree for the relation is 0.90.

4.2 Contextual Relevance

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 also holds when dealing with image analysis problems. Since visual context is a difficult notion to grasp and capture [20], we restrict it herein to the notion of ontological context, defined in terms of the fuzzy ontologies presented in subsection 4.1.

```
<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. 3. Fuzzy relation representation: RDF reification.

Here the problems to be addressed include how to meaningfully re-adjust the membership degrees of segmented (and possibly merged) image regions, and how to use visual context to improve the performance of knowledge-assisted image analysis. Based on the mathematical background introduced in the previous subsections, we develop an algorithm used to re-adjust 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 concept $k \in C$ in the application-domain's ontology is stored together with its relationship degrees r_{kl} to every other related concept $l \in C$. To tackle cases in which more than one concept is related to multiple concepts, we use the term *context relevance* $cr_{dm}(k)$, 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 $motorsports$. 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.

An estimation of each concept's value is derived from the direct and indirect relationships between the concept and other concepts, using a *compatibility indicator* or distance metric. The ideal distance metric for two concepts is one that quantifies their semantic correlation. Depending on the nature of the domains under consideration, the best indicator could be either the max or the min operator. For the problem at hand, *beach* and *motorsports* domains, the max value is a meaningful measure of correlation for both. Fig. 4 presents a simple example. The concepts are: $motorsports$ (the *root element* - denoted as m), $asphalt$ (a), $grass$ (g) and car (c). Their relationships can be summarised as follows: 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}. The context relevance for concept a is given by:

$$cr_{dm}(a) = \max\{r_{am}, r_{ag}r_{gm}, r_{ac}r_{cm}, r_{ag}r_{cg}r_{cm}, r_{ac}r_{cg}r_{gm}\} \tag{14}$$

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

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 $np \in [0, 1]$.

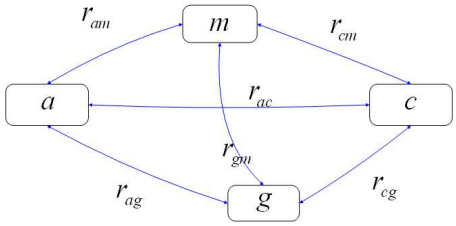


Fig. 4. Graph representation example - Compatibility indicator estimation.

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 contextual information in the form of its relations to all other concepts: $\{r_{kl} : l \in C, l \neq k\}$.
3. 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: $cr_k = max\{r_{kr^e}, \ldots, r_{km}\}$. We express the calculation of $\mu_a(k)$ with the recursive formula:

$$\mu_a^n(k) = \mu_a^{n-1}(k) - np(\mu_a^{n-1}(k) - cr_k) \tag{15}$$

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

$$\mu_a^n(k) = (1 - np)^n \cdot \mu_a^0(k) + (1 - (1 - np)^n) \cdot cr_k \tag{16}$$

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

In practice, typical values for n reside between 3 and 5. Interpretation of both equations (15) and (16) implies that the proposed contextual approach will favor confident degrees of membership for a region's concept over non-confident or misleading degrees of membership. It amplifies their differences, while diminishing confidence in clearly misleading concepts for each region. Further, based on the supplied ontological knowledge, it will clarify and solve ambiguities in cases of similar concepts or difficult-to-analyze regions.

The key step remaining is the definition of a meaningful normalization parameter np. When re-evaluating the confidence values, the ideal np is always defined with respect to the particular domain of knowledge and is the value that quantifies their semantic correlation to the domain. In our work we conducted a series of experiments on a training set of 120 images for both the beach and the motorsports application domains and selected the np value that resulted in the best overall evaluation score values for each domain. The proposed algorithm re-adjusts the initial degrees of membership, using semantics in the form of the contextual information residing in the constructed fuzzy ontology.

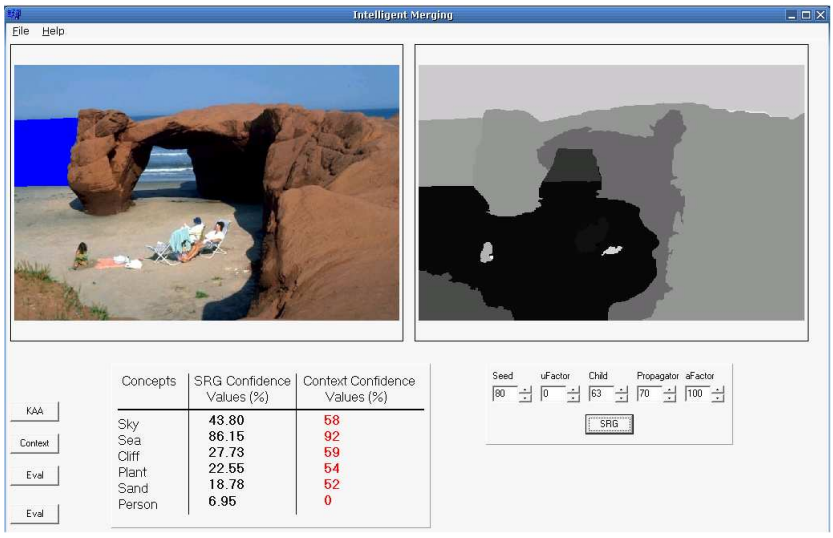


Fig. 5. Contextual experimental results for a beach image.

Fig. 5 presents indicative results for a beach image. Contextualization, which works on a per region basis, is applied after semantic region growing. In this example we have selected the unified sea region in the upper left part of the image (illustrated by an artificial 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 92.00%. The (irrelevant for this region) membership degree for person is extinguished, whereas degrees of membership for the rest of the possible beach concepts are slightly increased, due to the ontological knowledge relations that exist in the knowledge model.

5 Spatial Context and Optimization

5.1 Introduction

In this section, a semantic image analysis approach based on the incorporation of spatial-related contextual information in the analysis process and the formulation of the latter as a global optimization problem is presented. In particular, the examined image is spatially segmented and Support Vector Machines (SVMs) are subsequently employed for performing an initial association of every image region with a set of pre-defined high-level semantic concepts based solely on visual information. Then, a Genetic Algorithm (GA) is introduced for estimating a globally optimal region-concept assignment, taking into account the spatial context. Representation of the latter relies on fuzzy directional relations extraction.

5.2 Low-Level Visual Information Processing

Segmentation and Features Extraction

In order to perform the initial region-concept association procedure, the examined image has to be segmented into regions and suitable low-level descriptions have to be extracted for every resulting segment. In the current implementation, an extension of the Recursive Shortest Spanning Tree (RSST) algorithm has been used for segmenting the image. Output of this segmentation algorithm is a segmentation mask, where the created spatial regions $s_n,\ n = 1, ...N$, are likely to represent meaningful semantic objects. For every generated image segment the following MPEG-7 descriptors, discussed in Chapter 4, are extracted and form a *region feature vector*: *Scalable Color*, *Homogeneous Texture*, *Region Shape* and *Edge Histogram*.

Fuzzy Spatial Relations Extraction

In the present analysis framework, eight fuzzy directional relations are supported, namely *North* (N), *East* (E), *South* (S), *West* (W), *South-East* (SE), *South-West* (SW), *North-East* (NE) and *North-West* (NW). Their extraction builds on the principles of projection- and angle- based methodologies and consists of the following steps. First, a *reduced box* is computed from the *ground* region's (the region used as reference and is painted in dark grey in Fig. 6) MBR, so as to include the region in a more representative way. The computation of this *reduced box* is performed in terms of the MBR compactness value v, which is defined as the fraction of the region's area to the area of the respective MBR: If the initially computed v is below a threshold T, the ground region's MBR is reduced repeatedly until the desired threshold is satisfied. Then, eight cone-shaped regions are formed on top of this reduced box, as illustrated in Fig. 6, each corresponding to one of the defined directional relations. The percentage of the *figure* region (whose relative position is to be estimated and is painted in light grey in Fig. 6) points that are included in each of the cone-shaped regions determines the degree to which the corresponding directional relation is satisfied. After extensive experimentations, the value of the threshold T was set equal to *0.85*.

5.3 Initial Region-Concept Association

SVMs, which were discussed in Chapter 5, have been widely used in semantic image analysis tasks due to their reported generalization ability [21]. Under the proposed approach, SVMs are employed for performing an initial association of the computed image regions to one of the defined high-level semantic concepts based on the estimated region feature vector. An individual SVM is introduced for every defined concept c_l, $l = 1, ...L$, to detect the corresponding instances, and is trained under the '*one-against-all*' approach. Each SVM at

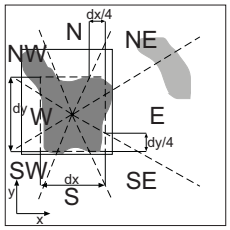


Fig. 6. Fuzzy directional relations definition

the evaluation stage returns for every segment a numerical value in the range $[0, 1]$ denoting the degree of confidence, h_{nl}^C, to which the corresponding region is assigned to the concept associated with the particular SVM. The degree of confidence is calculated according to the following equation:

$$h_{nl}^C = \frac{1}{1 + e^{-p \cdot z_{nl}}} \quad , \tag{17}$$

where z_{nl} is the distance of the input feature vector from the corresponding SVM's separating hyperplane and p is a slope parameter set experimentally. For every region, $argmax(h_{nl}^C)$ indicates its concept assignment, whereas $H_n^C = \{h_{nl}^C,\ l = 1, ...L\}$ constitutes its concept hypothesis set. It must be noted that any other classification algorithm can be adopted during this step, provided that a similar hypothesis set is estimated for every image region.

5.4 Final Region-Concept Association

Spatial Constraints Estimation

In this section, the procedure followed for estimating the values of the spatial relations (spatial-related contextual information) between all the defined high-level semantic concepts, as opposed to concepts themselves that are empirically determined, is described. Specifically, the aforementioned values are calculated according to the following learning approach:

Let R,

$$R = \{r_k,\ k = 1, ..., K\} = \{\ N,\ NW,\ NE,\ S,\ SW,\ SE,\ W,\ E\}, \tag{18}$$

denote the set of the supported spatial relations. Then, the degree to which region s_i satisfies relation r_k with respect to region s_j can be denoted as $I_{r_k}(s_i, s_j)$ and is estimated according to the procedure of Section 5.2. In order to acquire the contextual information, this function needs to be evaluated over a set of segmented images with ground truth annotations, that serves as a training set. For that purpose, an appropriate image set, $\mathcal{B}_{tr}$, is assembled. Then, using this training set the mean values, $I_{r_k mean}$, of I_{r_k} for every k over

all pairs of regions assigned to concepts (c_i, c_j), $i \neq j$, are estimated. These constitute the constraints input to the optimization problem which is solved by the genetic algorithm, as will be described in the sequel.

Spatial Constraints Verification Factor

The learnt fuzzy spatial relations, which are obtained as described in Section 5.4, serve as constraints denoting the "allowed" spatial topology of the supported concepts. In this section, the exploitation of these constraints is detailed. In particular, let $I_S\ (g_{ij},\ g_{pq})$ be defined as a function that receives values in the interval $[0, 1]$ and which returns the degree to which the spatial constraint between the g_{ij}, g_{pq} concept to region mappings is satisfied. To calculate this value the following procedure is followed: Initially, a normalized euclidean distance $d(g_{ij}, g_{pq})$ is calculated based on the following equation:

$$d(g_{ij}, g_{pq}) = \frac{\sqrt{\sum_{k=1}^{8}(I_{r_k mean}(c_j, c_q) - I_{r_k}(s_i, s_p))^2}}{\sqrt{8}} \quad , \tag{19}$$

which receives values in the interval $[0, 1]$. The function $I_S\ (g_{ij},\ g_{pq})$ is then defined as:

$$I_S\ (g_{ij},\ g_{pq}) \ = \ 1 - d(g_{ij}, g_{pq}) \tag{20}$$

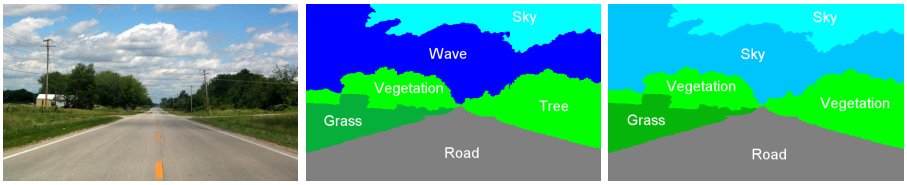
Implementation of Genetic Algorithm

As already described, the employed genetic algorithm realizes semantic image analysis as a global optimization problem, while taking into account both visual and spatial-related information. Under the proposed approach, each chromosome represents a possible solution. Consequently, the number of the genes comprising each chromosome equals the number N of the regions s_i produced by the segmentation algorithm and each gene assigns a supported concept to an image segment.

An initial population of 200 randomly generated chromosomes is employed. An appropriate *fitness function* is introduced to provide a quantitative measure of each solution fitness, i.e. to determine the degree to which each interpretation is plausible:

$$f(Q) \ = \ \lambda \cdot FS_{norm} \ + \ (1 - \lambda) \cdot SC_{norm} \ , \tag{21}$$

where Q denotes a particular chromosome, FS_{norm} refers to the degree of low-level descriptors matching, SC_{norm} stands for the degree of consistency with respect to the provided spatial knowledge, and variable λ, $\lambda \epsilon [0, 1]$, is introduced to adjust the impact of FS_{norm} and SC_{norm} on the final outcome.



Input image | Initial region-concept association | Region-concept association after application of the GA

Fig. 7. Indicative region-concept association results

The value of the latter is estimated according to an optimization procedure, as described in Section 5.4.

The values of SC_{norm} and FS_{norm} are computed as follows:

$$FS_{norm} = \frac{\sum_{i=1}^{N} I_M(g_{ij}) - I_{min}}{I_{max} - I_{min}}, \; SC_{norm} = \frac{\sum_{l=1}^{W} I_{S_l}(g_{ij}, g_{pq})}{W}, \quad (22)$$

where $I_M(g_{ij}) = h_{ij}^C$, $I_{min} = \sum_{i=1}^{N} \min_j I_M(g_{ij})$, $I_{max} = \sum_{i=1}^{N} max_j I_M(g_{ij})$, and W denotes the number of the constraints that had to be examined.

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

- Selection: the Tournament Selection Operator [8] with replacement is used for selecting a pair of chromosomes from the current generation to serve as parents for the generation of a new offspring.
- Crossover: uniform crossover with probability of 0.7 is used.
- Mutation: every gene of the processed offspring chromosome is likely to be mutated with probability of 0.008.

To ensure that chromosomes with high fitness will contribute to the next generation, the overlapping populations approach was adopted [17]. The above iterative procedure continues until the diversity of the current generation is equal to/less than 0.001 or the number of generations exceeds 50. In Fig. 7, indicative region-concept association results from the application of the proposed approach in outdoor images are presented.

Parameter Optimization

Since the selection of the value of parameter λ (Eq. 21) is of crucial importance in the behavior of the overall approach, its value is estimated according to a particular methodology. This methodology is also based on the use of a GA. Specifically, subject to the problem of concern is the computation of the value of parameter λ that leads to the highest correct concept association rate. For

that purpose, *Concept Accuracy*, CoA, is used as a quantitative performance measure and is defined as the fraction of the number of the correctly assigned concepts to the total number of image regions to be examined.

For optimizing parameter λ, each GA's chromosome Q represents a possible solution, i.e. a candidate λ value. Under the proposed approach, the number of genes of each chromosome is set equal to 5. The genes represent the binary coded value of parameter λ assigned to the respective chromosome, according to the following equation:

$$Q \;=\; [\; q_1 \; q_2 \; ...q_5 \;] \;\; where \;\; \sum_{i=1}^{5} q_i \cdot 2^{-i} \;=\; \lambda \tag{23}$$

where $q_i \;\epsilon\; \{0, 1\}$ represents the value of gene i. With respect to the corresponding GA's fitness function, the latter is defined as equal to the CoA metric already defined, where CoA is calculated over all images that are included in a validation set $\mathcal{B}_{val}$ (similar to the set $\mathcal{B}_{tr}$ defined in Section 5.4), after applying the GA of Section 5.4 with $\lambda = \sum_{i=1}^{5} q_i \cdot 2^{-i}$.

Regarding the GA's implementation details, an initial population of 100 randomly generated chromosomes is employed. New generations are successively produced based on the same evolution mechanism as described in Section 5.4. The differences are that the maximum number of generations is set equal to 30 and the probabilities of mutation and crossover are set equal to 0.4 and 0.2, respectively.

References

1. T. Adamek, N. O'Connor, and N. Murphy. Region-based segmentation of images using syntactic visual features. In *In Proc. of Workshop on Image Analysis for Multimedia Interactive Services*, Montreux, Switzerland, April 2005.
2. 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*, 34(2), March 2004.
3. G. Akrivas, M. Wallace, G. Andreou, G. Stamou, and S. Kollias. Context - sensitive semantic query expansion. In *In Proc. of the IEEE International Conference on Artificial Intelligence Systems*, Divnomorskoe, Russia, September 2002.
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. In *In Proc. of 5th International Workshop on Knowledge Markup and Semantic Annotation*, Galway, Ireland, November 2005.
5. F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and P.F. Patel-Schneider. *The Description Logic Hand-book: Theory, Implementation and Application*. Cambridge University Press, 2002.
6. S. Berretti, A. Del Bimbo, and E. Vicario. Efficient matching and indexing of graph models in content-based retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(12):1089–1105, December 2001.

7. S. Beucher and F. Meyer. *The Morphological Approach to Segmentation: The Watershed Transformation.* Marcel Dekker, NY, 1993.
8. D.E. Goldberg and K. Deb. A Comparative Analysis of Selection Schemes Used in Genetic Algorithms. *Urbana*, 51:61801–2996, 1991.
9. L. Hollink, G. Nguyen, G. Schreiber, J. Wielemaker, B. Wielinga, and M. Worring. Adding Spatial Semantics to Image Annotations. *Proc. of the 4th Int. Workshop on Knowledge Markup and Semantic Annotation at ISWC'04*, 2004.
10. MPEG ISO/IEC FDIS 15938-5 JTC1/SC29/WG11/M4242. Information technology - multimedia content description interface: Multimedia description schemes. Technical report, October 2001.
11. G. Klir. and B. Yuan. *Fuzzy Sets and Fuzzy Logic, Theory and Applications.* New Jersey, Prentice Hall, 1995.
12. J. Luo, M. Boutell, and C. Brown. Pictures are not taken in a vacuum. *Signal Processing Magazine, IEEE*, 23(2):101–114, 2006.
13. B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6):703–715, June 2001.
14. V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. Still image segmentation tools for object-based multimedia applications. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(4):701–725, June 2004.
15. V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. Segmentation of images and video. *Encyclopedia of Multimedia, B. Furht (Editor), Springer*, 2006.
16. C. Millet, I. Bloch, P. Hede, and P.A. Moellic. Using relative spatial relationships to improve individual region recognition. *Proc. 2nd Eur. Workshop Integration Knowledge, Semantics and Digital Media Technol*, pages 119–126, 2005.
17. M. Mitchel. An Introduction to Genetic Algorithms. *a Bradford Book, the MIT Press, Cambridge, Massachusetts*, 1996.
18. S. Miyamoto. *Fuzzy Sets in Information Retrieval and Cluster Analysis.* Kluwer Academic Publishers, 1990.
19. O.J. Morris, M.J. Lee, and A.G. Constantinides. Graph theory for image analysis: An approach based on the shortest spanning tree. *Institute of Electrical Engineering, pt. F*, 133(2):146–152, April 1986.
20. Ph. Mylonas and Y. Avrithis. Context modeling for multimedia analysis and use. In *In Proc. of 5th International and Interdisciplinary Conference on Modeling and Using Context*, Paris, France, July 2005.
21. G.T. Papadopoulos, V. Mezaris, I. Kompatsiaris, and MG Strintzis. Combining global and local information for knowledge-assisted image analysis and classification. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.
22. S. Staab and R. Studer. *Handbook on ontologies, international handbooks on information systems.* Heidelberg: Springer-Verlag, 2004.
23. W3C. Rdf. http://www.w3.org/RDF/.
24. W3C. Rdf reification. http://www.w3.org/TR/rdf-schema/.
25. W3C. Owl web ontology language reference. http://www.w3.org/TR/owl-ref/, February 2004.
26. Y. Wang, F. Makedon, J. Ford, L. Shen, and D. Goldin. Generating fuzzy semantic metadata describing spatial relations from images using the R-histogram. *Digital Libraries, Proc. of ACM/IEEE Conf. on*, pages 202–211, 2004.