Introducing Context and Reasoning in Visual Analysis: An Ontology-based Framework

Stamatia Dasiopoulou[1], Carsten Saathoff[2], Phivos Mylonas[3], Yannis Avrithis[3], Yiannis Kompatsiaris[1], and Steffen Staab[2]

[1] Multimedia Knowledge Group, Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece {dasiop,ikom}@iti.gr
[2] ISWeb - Information Systems and Semantic Web, University of Koblenz, Germany, {saathoff,staab}@uni-koblenz.de
[3] Image Video and Multimedia Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, Greece, {fmylonas,iavr}@image.ntua.gr

1 Introduction

The amount of multimedia content produced and made available is constantly growing, resulting in equally increasing needs in terms of efficient and effective ways to access it. Enabling smooth access at a level that meets user expectations and needs however to such content has been the holy grail in content-based retrieval for decades as it is intertwined with the so called *semantic gap* between the features that can be automatically extracted from content and the conveyed meaning as perceived by end users. Numerous efforts have been reported addressing a variety of domains and applications, and following diverse methodologies. Among the reported literature, knowledge-based approaches utilizing explicit prior knowledge constitute a popular choice. Customized as well as formal representations and corresponding reasoning services have been employed to enable analysis decouple from application-specific algorithmic implementations and benefit from the existence of prior knowledge about the examined problem. Apart from the followed knowledge representation formalism, knowledge-based approaches are further characterized by the types of knowledge employed (e.g., visual, spatial, temporal). Among these types of knowledge, contextual one is of particular interest as it can significantly assist in disambiguation and validation of the meaning extracted.

Following the Semantic Web (SW) vision of knowledge sharing and reuse through machine processable metadata, ontologies have been favored in recent knowledge-based image analysis efforts. In the works presented in (Little and Hunter 2004; Hollink, Little and Hunter 2005), ontologies have been used to represent concepts of the examined domain and their visual characteristics in terms of MPEG-7 descriptions. Ontologies have been also utilized in (Maillot and Thonnat 2005) to represent domain

knowledge, visual related one in terms of qualitative descriptions, and contextual with respect to image capturing conditions, while in (Dasiopoulou, Mezaris, Kompatsiaris, Papastathis and Strintzis 2005), ontologies combined with rules have been proposed to capture the processing steps required for the detection of the respective domain concepts. In the approaches presented in (Schober, Hermes and Herzog 2004) and (Neumann and Möller 2004), the inference services provided by DLs have been employed to perform scene interpretation tasks based on the corresponding domain ontologies definitions that link domain concepts with their visual characteristics.

Following the rationale of using ontologies to formally capture the semantics of the knowledge required for semantic image analysis, we proposed in this chapter an ontology-based framework for enhancing the descriptions resulting from typical image analysis through the exploitation of topological and visual context information. In the proposed framework, RDFS ontologies are used to capture not only the domain concepts of interest and their corresponding visual features, but also spatial and logical relations that constitute their context of appearance. Furthermore, the use of ontologies for structuring the content annotations (labels) produced, ensures smooth communication among the different modules involved and facilitates interoperability with respect to future extensions with additional modules.

As illustrated in Fig. 1, under the proposed framework the output of image analysis is considered as the initial labelling on top of which the developed context analysis and constraint reasoning modules are applied. The only assumption made with respect to image analysis is that the produced labels are associated with a degree of confidence and that more than one labels may exist for a given image region. It is easy to see that this assumption is not restricting but instead reflects the actual case in image analysis where due to the inherent ambiguity no certain descriptions can always be acquired and different analysis modules can be used, thus resulting in multiple labellings. Within such setting, the contextual information in terms of context of appearance is utilized to refine the initially produced degrees of confidence and then the constraint reasoning is applied to impose consistency on the labelling with respect to the provided domain topology.

The rest of the chapter is organized as follows. Section 2 presents relevant work in terms of utilizing context information and constraint solving approaches in semantic image analysis. In the following section the proposed framework architecture is described, including the specifications and design of the ontology infrastructure (Section 3.1) and the visual analysis component description (Section 3.2) respectively. Section 4 details the modelling and ontological representation of context of appearance, and presents the methodology for readjusting the initial degrees of confidence. Section 5.1 describes the application of constraint reasoning in image labelling. Experimental results and evaluation of the proposed framework are presented in Section 6, while Section 7 concludes the chapter with future directions.

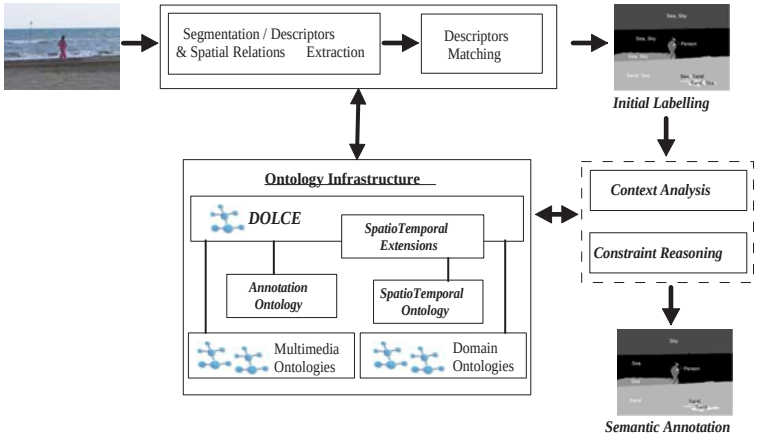

Fig. 1. Ontology-based framework introducing context and constraint reasoning in image analysis.

2 Relevant Work

2.1 Context in Multimedia Analysis

In semantic content-based image search and retrieval, research has shifted beyond low-level color, texture, and shape features in pursuit of more effective methods of content access at the level of the meaning conveyed. Towards this goal, context plays a significant role as it allows to enhance performance by exploiting the semantic correlation between the considered concepts. The added value of using context in image analysis becomes more apparent when considering the number of analysis errors that often occur because of the similarities in visual features such as color, texture, edge characteristics and so on, of the concepts considered.

A number of interesting efforts have been reported including among other the impact of exploiting co-occurrence information for the detection of natural objects in outdoor images (Vailaya and Jain 2000; Naphade, Kozintsev and Huang 2002). In (Luo, Singhal, and Zhu 2002), a spatial context-aware object detection system is presented that combines the output of individual object detectors into a composite belief vector for the objects potentially present in an image. In (Murphy, Torralba, and Freeman 2003) scene context is proposed as an extra source of global information to assist in resolving local ambiguities, while in (Boutel 2006) three types of context are explored for the scene classification problem: *spatial*, *temporal* and *image capture condition context* in the form of camera parameters, also examined in (Boutell and Luo 2005). Context information in terms of a combination of a region index and a presence vector has been proposed in (Le Saux and Amato 2004) for scene classification.

However, so far the investigated analysis approaches tend not to take into account contextual knowledge, neither with respect to the entire scene nor the individual objects themselves. This constitutes a major limitation as the notion of visual context is

able to aid in the direction of natural object detection methodologies, simulating the human approach to similar problems, one important trait of which is that the examination of all the objects in the scene before making a final decision on the identity of the individual objects. The use of visual context knowledge as described in Section 4 of the current manuscript provides the means to introduce and benefit from contextual knowledge, while additionally providing a formal model of its semantics.

2.2 Constraint Reasoning in Image Understanding

Constraint reasoning has a long history, starting with the system SKETCHPAD (Sutherland 1963) in the early 60's. The *Scene Labelling Problem* (Waltz 1975) followed later, formalized the notion of constraints in order to solve the problem of deriving a 3D-interpretation of 2D line drawings. Haralick and Shapiro formulated this problem even more generally as the labelling of image segments based on automatic low-level processing techniques (Haralick and Shapiro 1979). However, this original work was mainly formal, introducing the consistent labelling problem as a general set of problems, while in the approach proposed in thsi chapter we provide a concrete instantiation of the scene labelling problem, deployed in a real application setting. Only a few other approaches exist that employ constraint reasoning to introduce explicit knowledge about spatial arrangements of real life objects into the image interpretation process. We will shortly discuss some of them in the following.

In (Kolbe 1998), constraint reasoning techniques is employed for the identification of objects in aerial images. In Kolbe's work, one main aspect is the handling of over-constrained problems. An over constraint problem is a constraint solving problem (CSP) in which not all constraints can be satisfied in parallel. In traditional constraint reasoning this would mean that no solutions exists and the problem is consequently unsolvable. Several techniques were proposed to solve such over-constrained problems, providing solutions that are close to optimal. Kolbe specifically introduces a solving technique based on a information theory-based evaluation measure. In addition, Kolbe uses specialized constraints between the image parts, that are hardly applicable to more generic domains that we consider.

In (Hotz and Neumann 2005) a configuration system is adopted to provide high-level scene interpretations. The system is evaluated on so-called table laying scenes, using well-defined domain models based on the spatial arrangements of the concepts found within the given domain. The underlying interpretation of the spatial knowledge is also based on the notion of constraints on variable assignments, although the terminology of constraint reasoning is not used. The whole approach does not focus solely on the application of spatial knowledge, but also on the inference of higher level knowledge and the scene-specific interpretation of the image. However, again the problem is extremely specific and relies on very well defined domain models that are hardly to exist for domains like holiday or family images.

Finally, an interesting approach is presented in (Srihari and Zhang 2000), where images are annotated semi-automatically and a user can manually prune the search space by specifying hints such as "A L-shaped building in the upper left corner." A constraint reasoner is employed to enforce the user hints. Obviously, this approach uses

the constraints in an ad-hoc manner, and not as a domain model, such as employed in our architecture.

3 Ontology-based Visual Analysis Framework

The proposed ontology-based framework aims to serve as a generic, easy to extend knowledge-based framework for performing semantic image analysis through the refinement and consistency checking of initially produced labellings. The intended usage purpose imposes certain requirements with respect to the knowledge infrastructure that constitutes the framework's backbone as well as the developed context analysis and constraint reasoning methodology.

A first requirement refers to the need for smooth communication among the involved modules while preserving the intended semantics. This practically means that the initially produced labellings, possibly by independent analysis modules, need to be correctly interpreted by the modules involved in the process of their further refinement, namely context analysis and constraint reasoning in our application scenario. Given that image analysis and labelling relates both to domain-specific aspects, i.e., the conceptualization of examined domain concepts and relations, as well as media related ones, i.e., the structure of the labelled image, the corresponding knowledge infrastructure needs to capture the knowledge of both aspects in an unambiguous, machine processable way. Following an ontology-based knowledge representation using semantic web technologies meets such requirements since well-defined shareable semantics are ensured.

Another important requirement relates to the need for enabling extensibility in terms of incorporating not only additional analysis modules that are built in compliance with the existing knowledge modelling, but also already existing modules that adhere to possibly different conceptualizations. To enable a smooth harmonization between such different domain or media related conceptualizations a reference point is needed so that the involved modules intended meanings, i.e., ontological commitments, can be disambiguated and correctly aligned. The use of a core ontology through the axiomatization it provides enables avoiding terminological and conceptual ambiguities. Furthermore, a significant aspect present not only in image but multimedia analysis in general is the uncertainty introduced in the analysis and interpretation process. Audiovisual information provides cues, in terms of either low or intermediate level features, for the high-level descriptions that represent the conveyed meaning, not evidences. Thus, a knowledge-based framework intended to serve as a generic applicable approach to image labelling refinement needs to provide support for representing and handling such uncertainty issues.

The aforementioned knowledge structure related requirements reflect on the context analysis and constraint reasoning specifications: they need to be able to read and manipulate such knowledge, as well as output their results correspondingly. Furthermore, another principle requirement underlying the latter modules constitutes in the need to be not application oriented but able to perform on a generic basis instead.

This practically means that both modules specifications and implementation need to be assumption free with respect to the employed analysis modules implementation.

In the following we briefly present the adopted ontology-based knowledge infrastructure and the analysis approach employed to produce the initial labellings. As will be shown in the sequel, a generic approach has been followed with respect to the employed analysis approach, a choice justified by the intended scope of the proposed context analysis and constraint reasoning framework, i.e., no dedicated analysis analysis modules with the highest possible performance are needed for the proposed framework to be of value.

3.1 Ontology Infrastructure

In the following we briefly describe the different ontologies constituting the employed knowledge infrastructure. For further details the reader is referred to (Bloehdorn, Petridis, Saathoff, Simou, Tzouvaras, Avrithis, Handschuh, Kompatsiaris, Staab, and Strintzis 2005).

Core Ontology

The role of the core ontology in this framework is threefold: i) to serve as a bridge between existing ontologies to which the different modules may adhere to, ii) to serve as a starting point for the construction of new ontologies, and iii) to provide a reference point for comparisons among different ontological approaches. In our framework, we utilize DOLCE (Gangemi, Guarino, Masolo, Oltramari and Schneider 2002), which was explicitly designed as a core ontology. As DOLCE is minimal in the sense it includes only the most reusable and widely applicable upper-level categories, rigorous in terms of axiomatization, as well as extensively researched and documented.

Spatio-Temporal Extensions Ontology

In a separate ontology we have extended the dolce:Region concept branch of DOLCE to accommodate topological and directional relations between regions. Directional spatial relations describe how visual segments are placed and relate to each other in 2D or 3D space (e.g., left and above), while topological spatial relations describe how the spatial boundaries of the segments relate (e.g., touches and overlaps). In a similar way, temporal relations have been introduced.

Visual Descriptor Ontology

The Visual Descriptor Ontology (VDO) models concepts and properties that describe visual characteristics of concepts. VDO follows the MPEG-7 visual descriptors specification (MPEG-7 Visual Part 2001), with some modification so as to translate the XML Schema and datatype definitions into a valid RDFS representation.

Multimedia Structure Ontology

The Multimedia Structure Ontology (MSO) models basic multimedia entities from the MPEG-7 Multimedia Description Scheme (MPEG-7 Multimedia Description Schemes 2001). More specifically, MSO covers the five MPEG-7 multimedia content types, i.e., Image, Video, Audio, Audiovisual and Multimedia, and their corresponding segment and decomposition relation types.

Annotation Ontology

The Annotation Ontology (AO) provides the schema for linking multimedia content items to the corresponding semantic descriptions, i.e., for linking image regions to domain concept and relation labels. Furthermore, it is the AO that models the uncertainty with respect to the extracted labellings and allows to associate a degree of confidence to each label produced by the analysis.

Domain Ontology

In the presented multimedia annotation framework, the domain ontologies are meant to model the conveyed content layer of multimedia content with respect to specific real-world domains, such as sports events like tennis. They serve a dual role: i) they provide the vocabulary to be used in the produced annotations, and ii) they provide the domain conceptualization utilized during retrieval. As aforementioned, each domain ontologies is explicitly aligned to the DOLCE core ontology, ensuring thereby interoperability between different domain ontologies possibly used by different analysis modules.

3.2 Visual Analysis

The overall architecture of the analysis is illustrated in Fig. 2. Initial sets of graded hypotheses, i.e. set of labels with associated degrees of confidence, are generated for each of the examined image segment utilizing the provided domain knowledge prototypical visual description instances. The retrieval of the corresponding concept prototypes and their descriptor instances is performed using the OntoBroker engine. OntoBroker needs to load the domain ontologies where the high-level concepts are defined, the Visual Descriptor Ontology (VDO) where the low-level visual descriptors are present, and the prototype instance files that include the actual knowledge base and provide the linking of domain concepts with descriptor instances. These instance files are created using the M-OntoMat-Annotizer tool.

The analysis process starts with a pre-processing step where region-based segmentation extracts the atom-regions of the examined visual content. Subsequently, the dominant color, homogeneous texture and region shape descriptors are extracted for each region together with the spatial relations (such as above, below, is-included-in) between adjacent regions. The next analysis step is to compute matching distances

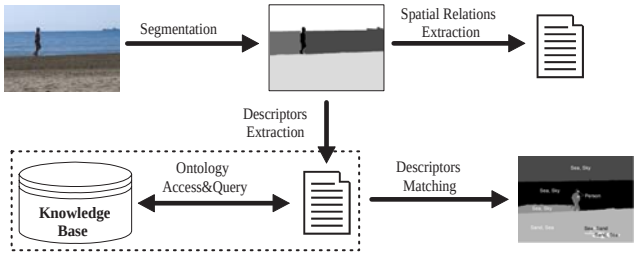

Fig. 2. Ontology-based framework introducing context and constraint reasoning in image analysis.

value between each one of the atom-regions and the prototype instances of all concepts in the domain ontology. This matching distance is evaluated by means of low-level visual descriptors. The generated metadata expresses the structure and semantics of the analyzed content, i.e. a number of segments or shots accompanied with the confidence values of the supported domain concepts. Additionally, the extracted spatial relations of each atom-region are also included. The extracted metadata are in RDFS format following the definitions of the Multimedia Structure Ontology (MSO) and the Annotation Ontology.

4 Context Analysis

4.1 Ontology-Based Contextual Knowledge Representation

In order to design a robust ontology-based contextual knowledge representation, we use a set of concepts and relations between them, as the basic elements towards semantic, as well as low-level, interpretation of context within the present research effort. Amongst all possible ways to describe the domain ontologies one can be formalized as follows:

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

where O is a domain ontology, C is a subset of the set of concepts described by the domain ontology, p and q are two concepts $p, q \in C$ and, R_{pq} is a possible semantic relation amongst these concepts. In general, semantic relations describe specific kinds of links or relationships between any two concepts. In the crisp case, a semantic relation either relates ($R_{pq} = 1$) or does not relate ($R_{pq} = 0$) a pair of concepts p, q with each other.

In addition, for a knowledge model to be highly descriptive, it must contain a large number of distinct and diverse relations among its concepts. A major side effect of this approach is the fact that available information will then be scattered among them, making each one of them inadequate to describe a context in a meaningful way. Consequently, relations need to be combined to provide a view of the knowledge that suffices for context definition and estimation. In this work we utilize three types of relations,

whose semantics are defined in the MPEG-7 standard, namely the *specialization* relation Sp, the *part* relation P and the *property* relation Pr.

The last point to consider when designing such a knowledge model is the fact that real-life data often differ from research data. Real-life information is in principal governed by uncertainty and fuzziness, thus herein its modelling is based on *fuzzy* relations. For the problem at hand, the above set of commonly encountered crisp relations can be modelled as fuzzy relations and can be combined for the generation of a meaningful fuzzy taxonomic relation, which will assist in the determination of context. Consequently, to tackle such complex types of relations we propose the following "fuzzification" of the previous domain ontology definition:

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

where O_F defines a "fuzzified" domain ontology, C is again a subset of all possible concepts it describes and r_{pq} denotes a fuzzy semantic 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 (Klir and Yuan 1995), a fuzzy set F on C is described by a membership function $\mu_F : C \rightarrow [0,1]$. We may describe the fuzzy set using the widely applied sum notation (Miyamoto 1990):

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

where $n = |C|$ is the cardinality of 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 Klir et al., a fuzzy relation on C is a function $r_{pq} : C \times C \rightarrow [0,1]$ and its inverse relation is defined as $r_{pq}^{-1} = r_{qp}$. Based on the relations r_{pq} and, for the purpose of image analysis, we construct the following relation T with use of the corresponding set of fuzzy relations Sp, P and Pr:

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

Based on the roles and semantic interpretations of Sp, P and Pr, as they are defined in the MPEG-7 standard (MPEG-7 Multimedia Description Schemes 2001), it is easy to see that equation (4) 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 set of the above relations is either defined explicitly in the domain ontology or is considered to be a superset of the set defined in the latter. Most commonly encountered, a domain ontology includes some relations between its concepts, that are all of the *SubclassOf* type and consequently, we extend it by defining additional semantic relations. The transitive closure relation extension Tr^t is required in both cases, in order for T to be taxonomic, as the union of transitive relations is not necessarily transitive, as discussed in (Akrivas, Wallace, Stamou, and Kollias 2002).

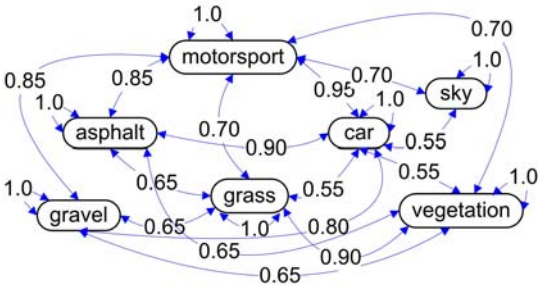

Fig. 3. Graph representation example - motorsports domain

The representation of this concept-centric contextual knowledge model follows the Resource Description Framework (RDF) standard (Becket 2004) proposed in the context of the Semantic Web. RDF is the framework in which Semantic Web metadata statements can be expressed and represented as graphs. Relation T can be visualized as a graph, in which every node represents a concept and each edge between two nodes constitutes a contextual relation between the respective concepts. Additionally each edge has an associated membership degree, which represents the fuzziness within the context model. A sample graph derived from the motorsports domain is depicted in Fig. 3.

Representing the graph in RDF is a straight forward task, since RDF structure itself is based on a similar graph model. Additionally, the *reification* technique (Brickley and Guha 2004) was used in order to achieve the desired expressiveness and obtain the enhanced functionality introduced by fuzziness. Representing the membership degree associated with each relation is carried out by making a statement about the statement, which contains the degree information. Representing fuzziness with such reified statements is a novel but acceptable way, since the reified statement should not be asserted automatically. For instance, having a statement such as: "*BeachScene part Sky*", which means that sky is part of a beach scene, and a membership degree of 0.75 for this statement, does obviously not entail that sky is always a part of a beach scene. A small illustrative example is provided in the following Table 1 for an instance of the specialization relation Sp. As defined in the MPEG-7 standard $Sp\,(x, y) > 0$ means 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. Additionally, the proposed reification process introduces a statement about the former statement on the *specializationOf* resource, by stating that 0.90 is the membership degree to this relation.

4.2 Visual Context Analysis

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 multi-

Table 1. Fuzzy relation representation: RDF reification.

<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:specializationOf rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.90</context:specializationOf>
</rdf:Description>

media analysis problems. Since visual context is acknowledged to be a difficult notion to grasp and capture (Mylonas and Avrithis 2005), we restrict it herein to the notion of ontological context. The latter is defined on the "fuzzified" version of traditional ontologies presented in 4.1.

In a more formal manner, the problem that this work attempts to address is summarized in the following statement: the visual context analysis algorithm readjusts in a meaningful way the initial label confidence values produced by the prior steps of multimedia analysis, described in detail in Section 3.2. In designing such an algorithm, contextual information residing in the aforementioned domain ontology is utilized. In general, the notion of context is strongly related to the notion of ontologies since an ontology can be seen as an attempt for modelling real-world (fuzzy) entities and context determines the intended meaning of each concept, i.e. a concept used in different context may have different meanings. In this section the problems to be addressed include how to meaningfully readjust the initial membership degrees and how to use visual context to influence the overall results of knowledge-assisted image analysis towards higher performance.

Based on the mathematical background described in detail in the previous subsections, we introduce the algorithm used to readjust the degree of membership $\mu_a(c)$ of each concept c in the fuzzy set of candidate labels $L_a = \sum_{i=1}^{|C|} c_i/\mu_a(c_i)$ associated to a region a of an image in an image scene and extracted in the process described in Section 3.2. Each specific concept $k \in C$ present in the application-domain's ontology is stored together with its relationship degrees r_{kl} to any other related concept $l \in C$.

Another important point to consider is the fact that each concept has a different probability to appear in the scene. A flat context model (i.e., relating concepts only to the respective scene type) would not be sufficient in this case. We model a more detailed graph where ideally concepts are all related to each other, implying that the graph relations used are in fact transitive. As can be observed in Fig. 3, every concept participating in the contextualized ontology has at least one link to the root element. Additional degrees of confidence exist between any possible connections of nodes in the graph, whereas the root beach element could be related either directly or indirectly with any other concept. To tackle cases that more than one concept is related to multiple concepts, the term context relevance $cr_{dm}(k)$ is introduced, which refers to the overall relevance of concept k to the root element characterizing each domain dm. For instance the root element of beach and motorsports domains are concepts and

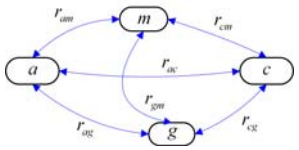

Fig. 4. Graph representation example - Compatibility indicator estimation.

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 provided in the domain ontology, the best indicator could be selected using the *max* or the *min* operator, respectively. Of course the ideal distance metric for two concepts is again one that quantifies their semantic correlation. For the problem at hand the *max* value is a meaningful measure of correlation for both of them. A simplified example derived again from the motorsports domain ontology, assuming that the only available concepts are $motorsports$ (the root element - denoted as m), $asphalt$ (a), $grass$ (g) and car (c) is presented in Fig. 4 and summarized in the following: let concept a be related to concepts m, g and c directly with: r_{am}, r_{ag} and r_{ac}, while concept g is related to concept m with r_{gm} and concept c is related to concept m with r_{cm}. Additionally, c is related to g with r_{cg}. Then, we calculate the value for $cr_{dm}(a)$:

$$cr_{dm}(a) = \max\{r_{am}, r_{ag}r_{gm}, r_{ac}r_{cm}, r_{ag}r_{cg}r_{cm}, r_{ac}r_{cg}r_{gm}\} \tag{5}$$

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 $np \in [0, 1]$.
2. For each concept k in the fuzzy set L_a associated to a region in a scene with a degree of membership $\mu_a(k)$, obtain the particular contextual information in the form of its relations to the set of any other concepts: $\{r_{kl} : l \in C, l \neq k\}$.
3. Calculate the new degree of membership $\mu_a(k)$ associated to region, 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:

$$\mu_a^n(k) = \mu_a^{n-1}(k) - np(\mu_a^{n-1}(k) - cr_k) \tag{6}$$

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

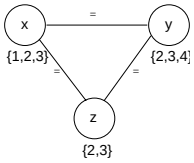

Fig. 5. A simple constraint satisfaction problem.

$$\mu_a^n(k) = (1 - np)^n \cdot \mu_a^0(k) + (1 - (1 - np)^n) \cdot cr_k \tag{7}$$

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

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

Key point in this approach remains the definition of a meaningful normalization parameter np. When re-evaluating this value, the ideal np is always defined with respect to the particular domain of knowledge and is the one that quantifies its semantic correlation to the domain. Application of a series of experiments on a training set of images for every application domain results in the definition of a np corresponding to the best overall evaluation score values for each domain. Thus, the proposed algorithm readjusts in a meaningful manner the initial degrees of membership, utilizing semantics in the form of the contextual information that resides in the constructed "fuzzified"ontology.

5 Constraint Reasoning to Eliminate Ambiguities in Labelled Images

5.1 Constraint Satisfaction Problems

Informally, a constraint satisfaction problem (CSP) consists of a number of variables and a number of constraints. A variable is defined by its domain, i.e. the set of values that can be assigned to the variable, and a constraint relates several variables and thereby restricts the legal assignments of values to each of the involved variables. *Constraint Reasoning* is the process of computing a solution to the given CSP, i.e. an assignment of values to the variables that satisfy all the given constraints on the variable.

In Fig. 5 a simple CSP is depicted, containing three variables x, y, and z and three constraints. The domains of x, y and z are $D(x) = \{1, 2, 3\}$, $D(y) = \{2, 3, 4\}$ and $D(z) = \{2, 3\}$. The constraints are $x = y$, $x = z$ and $y = z$, so that in a solution to the problem, the values of x, y and z must be equal.

Formally, a CSP consists of a set of variables $V = \{v_1, \ldots, v_k\}$ and a set of constraints $C = \{c_1, \ldots, c_l\}$. Each variable v_i has an associated domain $D(v_i) = \{l_1, \ldots, l_m\}$, which contains all values that can be assigned to v_i. Each constraint c_j is a relation on the domains of a set of variables, $v_1, \ldots, v_r \in V$, such that a constraint c_j is defined as $c_j \subseteq D(v_1) \times \cdots \times D(v_r)$. The constraint is said to be solved, iff both and $c_j = D(v_1) \times \cdots \times D(v_r)$ and c_j is non-empty. A CSP is solved, iff both all of its constraints are solved and no domain is empty, and failed, iff it contains either an empty domain or an empty constraint.

A variety of techniques have been proposed to solve constraint satisfaction problem, and they are usually collected under the name *Constraint Reasoning*. One can distinguish between two major types of solving techniques: consistency techniques and search methods. Consistency techniques try to simplify subproblems of a given CSP. However, a CSP that is locally consistent, i.e. where each relevant subproblem is consistent, is not necessarily (and in fact usually not) globally consistent. As an example consider *arc consistency*. Arc consistency only considers one constraint at a time. The constraint is said to be arc consistent, if for each assignment of a domain value to a variable of the constraint, assignments to all other related variables exist, that satisfy the constraint. This variable is said to have support in the other domains. A CSP is arc consistent if each of it's constraints are arc consistent.

Now, in the example of Fig. 5, the domain of x, y and z would all be reduced to $\{2, 3\}$ by an arc consistency algorithm. One can easily verify this, since an assignment of 1 to x would in every case violate the constraint $x = y$, since 1 is not a member of $D(y)$, and the same is true for an assignment $y = 4$, which neither has support in $D(x)$, nor in $D(z)$.

Local consistency can remove values from the domains of variables that will never take part in a solution. This can already be useful in some scenarios, but usually one searches for a concrete solution to a given CSP, i.e an unique assignment of values to variables that satisfy all the given constraints. As we can see from the example, an arc consistent CSP does not provide this solution directly. Obviously, assigning an arbitrary value from the remaining domains will not yield a valid solution. For instance, the assignment $x = 2, y = 2, z = 3$ only uses values from the arc consistent domains, but it is not a solution.

Therefore, in order to compute a concrete solution, search techniques are employed, such as backtracking. Often local consistency checks and search are integrated in hybrid algorithms, that prune the search space during search using local consistency notions and thus perform better. However, solving CSPs efficiently is highly problem specific, and a method that performs well for a specific problem, might have a much worse performance in another problem.

We will not further elaborate on local consistency notions and search techniques, since they are out of the scope of the paper. We assume that standard methods are employed to solve the constraint satisfaction problems we generate and that runtime performance is of lower priority. In general a good introduction to constraint reasoning is given in (Apt 2003). An overview of recent research in the field of constraint reasoning can be found in the survey presented in (Bartak 1999).

5.2 Image Labelling as a Constraint Satisfaction Problem

In order to disambiguate the region labels using a constraint reasoning approach, we have to

1. represent the employed knowledge as constraints, and
2. transform a segmented image into a CSP.

As we already mentioned in Section 3, spatial relations provide an important means to interpret images and to disambiguate region labels. Although heuristic, they give very valuable hints on what kind of object is depicted in a specific location. So, one would never expect a car depicted in the sky, or in the context of our framework, one would not expect the sky to be depicted below the sea in a beach image. Obviously, in order to use spatial knowledge for this kind of multimedia reasoning, the core elements are the spatial relations between the regions and the knowledge about the expected spatial arrangements of objects (i.e. labels) in a given domain.

It is obvious that, projected on the terminology of CSPs, the regions will become variables of the resulting CSP, and that the spatial relations will be modelled as constraints on those variables. In the following section we will first discuss how to define spatial constraints, and then, in the subsequent section, introduce the transformation of a initially labelled image into a CSP.

Spatial Constraints

The purpose of a spatial constraint is to reduce the number of labellings for a number of segments that are arranged in a specific spatial relationship. In other words, if a segment is above another segment, we want to make sure, that the lower segment only gets the label *Sky* if the upper one has a compatible label, such as *Sky* or *Cloud*. We will therefore define for each *spatial relation* that we want to consider a corresponding *spatial constraint type* that encodes the valid labellings as tuples of allowed labels. We will also call this set of tuples the domain of the constraint type. The concrete *spatial constraint* that is instantiated between a set of variables will then be formed by the intersection of the constraint type domain and the cross-product of the relevant variable domains.

Let SR now be the set of spatial relations under consideration and $r_t \in SR$ be a spatial relation of type t. Further, O is the set of all possible labels of a given application domain. We then define the domain of a spatial constraint type t to be $D(t) \subseteq O^n$, with n being the arity of the spatial relation. Obviously, each tuple in the domain of the constraint type, is supposed to be a valid arrangement of labels for the spatial relation of type t.

Now, let $V := \{v_1, \ldots, v_n\}$ be a set of variables related by a spatial relation $r_t \in SR$, and $D(t)$ the corresponding domain for the spatial relation. A constraint c^t_V of type t on the set of variables V is now defined as $c^t_V := D(t) \cap (D(v_1) \times \cdots \times D(v_n))$. Apparently, c^t_V now is a relation on the variable domains containing only those tuples that are allowed for the spatial relation r_t.

Transformation

In order to describe the transformation of an initially segmented and labelled image, we will shortly introduce some formal notions. Let a labelled image be a tuple $I = (S, SR)$, where S is the set of segments produced by the initial segmentation and SR is the set of spatial relations extracted by the spatial extraction module. For each segment $s \in S$ the hypothesis set of initial labels is denoted as $ls(s)$. The set of all possible labels is named O and $ls(s) \in O$ must hold. Each spatial relationship $r_t \in SR$ is of type t and has an associated domain of $D(t)$.

Transforming a labelled image into a CSP is now a straight forward process. For each segment a variable is created and a corresponding constraint is added for each spatial relation extracted. Obviously, the hypotheses sets become the domains of the variables. Currently we only consider two types of spatial relations: relative and absolute. Relative spatial relations are binary and derived from spatial relations that describe the relative position of one segment with respect to another one, such as *contained-in* or *above-of*. Absolute spatial constraints are derived from the absolute positions of segments on the image, like *above-all* and which are apparently unary constraints.

Let $I = (S, SR)$ be an labelled image as introduced above, then the algorithm to transform I into a corresponding CSP is as follows:

1. For each segment $s \in S$ create a variable v^s.
2. Set for newly created variable v^s the domain to $D(v^s) = ls(s)$.
3. Let SR be the set of all spatial relations defined in the domain knowledge, then
   a) add for each spatial relation $r_t \in SR$ between a number of segments $s_1, \ldots, s_n \in S$ a constraint $c^t_{\{v_1,\ldots,v_n\}}$ to the CSP, where $v_1, \ldots, v_n$ are the variables created from $s_1, \ldots, s_n$.

The result is a CSP conforming to what was introduced in Section 5.1. Standard constraint reasoning techniques can be used to solve the CSP, and because of the finiteness of the problem, all solutions can be computed. The latter property is quite useful, since the degree of confidence produced during the initial labelling, which is currently not employed during the constraint reasoning directly, can afterwards be used to rank the solutions according to the labels degrees. If only one solution would be computed, one would have to accept the first one found.

An example is depicted in Fig. 6, where the input image, the initial set of hypotheses, the corresponding initial labelling, and the labelling after the constraint reasoning application are depicted. We note that for the initial labelling, the labels with the highest score are kept for each segment. It is easy to see that two errors were made by the segment classification. The topmost segment was labelled as *Sea* instead of *Sky* and one of the small segments within the sand regions was labelled with *Sea*. After applying the constraint reasoning both erroneous labels have been corrected. For the topmost segment the absolute spatial relation *above-all* restricts the segment to the label *Sky* and the second wrong label was corrected using the *contained-in* constraint that does not allow a *Sand* segment to contain a *Sea* segment.

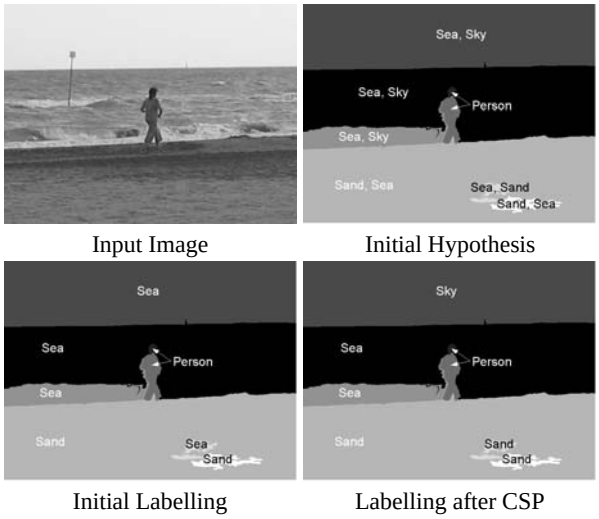

Fig. 6. Example of CSP application.

6 Experimental Results and Evaluation

In this section we present experimental results with respect to the enhancement achieved by the application of the proposed context analysis and constraint reasoning modules. For the experimentation, a set of 150 images from the beach domain has been assembled, 30 of which were used as the training set for estimating the parameter values required for context analysis. The resulting 120 images have first undergone the analysis steps presented in Section 3.2 in order to obtain the corresponding initial labellings. For each image region, the label with the highest degree of confidence from the respective hypotheses set is kept as the analysis output. In the sequel, the two modules have been applied. The context analysis module, exploiting the domain concepts associations and the information extracted through training, re-adjust the degrees of confidences towards more meaningful values. The constraint reasoner in the other hand applies its rules on the initial set of labels for each region resulting in the removal of those that were not consistent with the domain spatial topology. Again, to obtain the labelling generated by each module, the label with the highest degree are kept for each region respectively.

To overcome the difficulties and cost in defining generally accepted pre-annotated segmentation masks and avoid getting into segmentation evaluation process, a grid-based evaluation approach has been followed. This choice is justified by the given evaluation context as well, since, contrary to applications that require very accurate object boundaries detection, it allows certain tolerance for this kind of inaccuracies.

More specifically, in the proposed evaluation framework, ground truth construction and comparison against the examined annotations are both performed at block level. The grid size is selected with respect to the desired degree of evaluation precision: the

Table 2. Evaluation results for the beach domain

	Analysis			Context			CSP		
Concept	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Sky	0.77	0.69	0.73	0.85	0.69	0.76	0.78	0.91	0.84
Sea	0.66	0.59	0.62	0.64	0.67	0.65	0.73	0.53	0.62
Sand	0.75	0.94	0.84	0.81	0.94	0.87	0.85	0.97	0.90
Person	0.33	0.65	0.43	0.41	0.70	0.51	0.38	0.62	0.47

smaller the blocks size the increased the accuracy attained. To evaluate an annotation, the corresponding annotated mask is partitioned according to the selected grid size and the annotations within each block are compared to the ground truth ones.

To quantify the performance, we adopted the precision and recall metrics from the information retrieval (IR) field. For each domain concept *precision (p)* defines the proportion of correctly annotated segments *cf* over all the number of segments annotated with that concept *f*, while *recall (r)* is the proportion of correctly annotated segments over the number of segments depicting that concept in reality *c*. To determine the overall performance per concept, all *c*, *f* and *cf* for each of the respective concepts are added up, and using the above formulas, overall precision and recall values are calculated. Additionally, the *F-measure* was used to obtain a single metric. The *F-measure* is the harmonic mean of precision and recall, i.e., $F = 2pr/(p + r))$ and contrary to arithmetic mean it gets large only if both precision and recall are large. In case a concept was not depicted in an image at all, all three values are set to 0, so that they do not influence the overall computation.

In the current experimentation, four concepts have been considered, namely *Sky*, *Sea*, *Sand* and *Person*. In Table 2 the precision, recall and F-measure are given for the examined test images with respect to sole analysis, and the application of the proposed context and constraint reasoning approaches respectively. From the obtained results, one easily notes that almost in all cases precision and recall improve. The actual percentage of the gained performance improvement differs with respect to the concept considered, as each concept bears less or more semantic information. For example, the lower improvement is observed with respect to the concept Person, as due to over- and under-segmentation phenomena and the effects from the transition from $2D$ to $3D$ region depicting a Person may appear almost in any topological arrangement with respect to the rest of the domain concepts considered. Similar for its visual context of appearance, with respect of course to the examined domain.

7 Conclusions and further discussions

In this chapter we have proposed an ontology-based framework for semantic image analysis through the refinement of initially produced labellings by means of explicit knowledge about context of appearance and spatial constraints of the examined domains. Following the proposed framework one can smoothly integrate independent

analysis modules benefiting from the knowledge sharing facilities provided by the use of ontologies and from the sole dependency of context analysis and constraint reasoning from the available knowledge that decouples them from the actual analysis. Consequently, the main contributions of the proposed framework summarize in the following: i) the introduction of context of appearance information in a formal ontological way, and ii) the adoption of a constraint problem solving methodology within the image labelling domain, and iii) the applicability and extensibility of the presented framework to a variety of image analysis applications.

Future directions include further investigation of the proposed framework using more concepts, thereby making available additional knowledge, i.e., more spatial constrains and contextual associations. Similarly experimentation with alternative analysis modules or their combination would provide useful and concrete insight into the proposed framework contribution in real applications scenarios.

References

Akrivas, G. Wallace, M., Andreou, G., Stamou, G. and Kollias, S. (2002) *Context - Sensitive Semantic Query Expansion*. In: Proc. of the IEEE International Conference on Artificial Intelligence Systems (ICAIS), Divnomorskoe, Russia.

Apt, K. (2003) *Principles of Constraint Programming*. In: Cambridge University Press.

Bartak, R. (1999) *Constraint Programming: In Pursuit of the Holy Grail*. In: Proc. of Week of Doctoral Students (WDS99), pp. 555-564.

Becket, D. (2004) *RDF Schema Specification 1.0, W3C Recommendation 10 February 2004*, http://www.w3.org/TR/rdf-schema/.

Benitez, A., Zhong, D., Chang, S. and Smith, J. (2001) *MPEG-7 MDS content description tools and applications*. In: Proc. of International Conference on Computer Analysis of Images and Patterns (CAIP), Warsaw, Poland.

Brickley, D. and Guha, R.V. (2004) *RDF Schema Specification 1.0, W3C Recommendation 10 February 2004*, http://www.w3.org/TR/rdf-schema/.

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

Boutell, M. (2006) *Exploiting Context for Semantic Scene Classification*. In: Technical Report 894 (Ph.D. Thesis), University of Rochester, (https://urresearch.rochester.edu/retrieve/5932/tr894.pdf).

Boutell, M. and Luo, J. (2005) *Beyond pixels: Exploiting camera metadata for photo classification*. In: Pattern Recognition 38(6).

Dasiopoulou, S., Mezaris, V., Kompatsiaris, I., Papastathis, V.K., and Strintzis, M.G. (2005) *Knowledge-assisted semantic video object detection*. In: IEEE Trans. on Circuits and Systems for Video Technology, vol. 15, no 10, pp. 1210-1224.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. and Schneider, L. (2002) *Sweetening Ontologies with DOLCE*. In: Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Proceedings of the 13th International Conference on Knowledge Acquisition, Modeling and Management (EKAW), Siguenza, Spain.

Haralick, R.M. and Shapiro, L.G. (1979) *The Consistent Labeling Problem: Part I*. In: IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 1, pp. 173-184.

Hollink, L., Little, S. and Hunter, J. (2005) *Evaluating the application of semantic inferencing rules to image annotation*. In: K-CAP, pp. 91-98.

Hotz, L. and Neumann, B. (2005) *Scene Interpretation as a Configuration Task*. In: Künstliche Intelligenz, pp. 59-65.

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

Kolbe. T.H. (1998) *Constraints for Object Recognition in Aerial Images - Handling of Unobserved Features*. In: Lecture Notes in Computer Science, vol. 1520.

Le Saux, B., Amato, G. (2004) *Image classifiers for scene analysis*. In: International Conference on Computer Vision and Graphics (ICCVG), Warsaw, Poland.

Little, S. and Hunter, J. (2004) *Rules-By-Example - A Novel Approach to Semantic Indexing and Querying of Images*. In: International Semantic Web Conference (ISWC), pp. 534-548.

Luo, J., Singhal, A., and Zhu, W. (2002) *Natural Object Detection in Outdoor Scenes Based on Probabilistic Spatial Context Models*. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME).

Maillot, N. and Thonnat, M. (2005) *A Weakly Supervised Approach for Semantic Image Indexing and Retrieval*. In: CIVR, pp. 629-638.

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

MPEG-7 Visual Part (2001). ISO/IEC 15938-3 FCD Information Technology - Multimedia Content Description Interface - Part 3: Visual, March 2001, Singapore.

MPEG-7 Multimedia Descritpion Schemes (2001). ISO/IEC 15938-5 FCD Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Scemes, March 2001, Singapore.

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

Murphy, P., Torralba, A., and Freeman, W. (2003) *Using the forest to see the trees: a graphical model relating features, objects and scenes*. In: Adv. in Neural Information Processing Systems 16 (NIPS), Vancouver, BC, MIT Press.

Naphade, M., Kozintsev, I. and Huang, T.S. (2000)*A factor graph framework for semantic indexing and retrieval in video*. In: CIEEE Trans. Circuits Syst. Video Techn., vol. 12, no 1, pp. 40-52.

Neumann, B. and Möller, R. (2004) *On Scene Interpretation with Description Logics*. In: Technical report FBI-B-257/04, University of Hamburg, Computer Science Department.

Schober, J.P, Hermes, T. and Herzog, O. (2004) *Content-based Image Retrieval by Ontology-based Object Recognition*. In: Proceedings of the KI-2004 Workshop on Applications of Description Logics (ADL), Ulm, Germany.

Sikora, T. (2001) *The MPEG-7 Visual standard for content description - an overview*. In: Special Issue on MPEG-7, IEEE Trans. on Circuits and Systems for Video Technology, 11/6:696-702, June.

Srihari, R.K. and Zhang, Z. (2000) *Show&Tell: A Semi-Automated Image Annotation System*. In: IEEE MultiMedia, vol. 7, no 3, pp. 63-71.

Sutherland, I.E. (1963), *Sketchpad: A Man-Machine Graphical Communication System*. In: Ph.D Thesis.

Vailaya, A. and Jain, A. (2000) *Detecting sky and vegetation in outdoor images*. In: Proc. SPIE, vol. 3972, January.

Waltz, D. (1975) *Understanding line drawings of scenes with shadows*. In: The Psychology of Computer Vision,McGraw-Hill, Winston, Patrick Henry, New York.