Chapter 4
Introducing Context and Reasoning in Visual Content Analysis: An Ontology-Based Framework

Stamatia Dasiopoulou, Carsten Saathoff, Phivos Mylonas, Yannis Avrithis, Yiannis Kompatsiaris, Steffen Staab, and Michael G. Strinztis

4.1 Introduction

The amount of multimedia content produced and made available on the World Wide Web, and in professional and, not least, personal collections, 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 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 extracted from such content through automatic analysis and the conveyed meaning as perceived by the end users. Numerous efforts towards more reliable and effective visual content analysis that target the extraction of user-oriented content descriptions have been reported, addressing a variety of domains and applications, and following diverse methodologies. Among the reported literature, knowledge-based approaches utilising explicit, a priori, knowledge constitute a popular choice aiming at analysis methods decoupled from application-specific implementations. Such knowledge may address various aspects including visual characteristics and numerical representations, topological knowledge about the examined domain, contextual knowledge, as well as knowledge driving the selection and execution of the processing steps required.

Among the different knowledge representations adopted in the reported literature, ontologies, being the key enabling technology of the Semantic Web (SW) vision for knowledge sharing and reuse through machine processable metadata, have been favoured in recent efforts. Indicative state-of-the-art approaches include, among others, the work presented in Little and Hunter (2004), and Hollink, Little and Hunter (2005), where ontologies have been used to represent objects of the

S. Dasiopoulou
Multimedia Knowledge Laboratory, Centre for Research and Technology Hellas, Informatics and Telematics Institute, Thermi-Thessaloniki, Greece
e-mail: dasiop@iti.gr

Y. Kompatsiaris, P. Hobson (eds.), *Semantic Multimedia and Ontologies,*

examined domain and their visual characteristics in terms of MPEG-7 descriptions, and the ontological framework employed in Maillot and Thonnat (2005) that employs domain knowledge, visual knowledge in terms of qualitative descriptions, and contextual knowledge with respect to image capturing conditions, for the purpose of object detection. Furthermore, in Dasiopoulou, Mezaris, Kompatsiaris, Papastathis and Strintzis (2005), ontologies combined with rules have been proposed to capture the processing steps required for object detection in video, while in the approaches presented in Schober, Hermes and Herzog (2004) and Neumann and Möller (2004), the inference services provided by description logics (DLs) have been employed over ontology definitions that link domain concepts and visual characteristics.

In this chapter, we propose an ontology-based framework for enhancing segment-level annotations resulting from typical image analysis, through the exploitation of visual context and topological information. The concepts (objects) of interest and their spatial topology are modelled in RDFS (Brickley and Guha 2004) ontologies, and through the use of reification, a fuzzy ontological representation is achieved, enabling the seamless integration of contextual knowledge. The formalisation of contextual information enables a first refinement of the input image analysis annotations utilising the semantic associations that characterise the context of appearance. For example, in an image from the beach domain, annotations corresponding to concepts such as *Sea* and *Sand* are favoured contrary to those referring to concepts such as *Mountain* and *Car*. The application of constraint reasoning brings further improvement, by ensuring the consistency of annotations, through the elimination of annotations violating the domain topology semantics, such as the case of the *Sky*-annotated segment on the left of the *Sea*-annotated segment in Fig. 4.1.

Thereby, as illustrated in Fig. 4.1, the image analysis part is treated as a black box that provides initial annotations on top of which the proposed context analysis and constraint reasoning modules perform to provide for more reliable content descriptions. The only requirement with respect to the image analysis is that the produced annotations come with an associated degree of confidence. It is easy to see that such a requirement is not restricting but instead reflects the actual case in image analysis, where due to the inherent ambiguity, the similarities shared among different objects, and the different appearances an object may have, it is hardly possible to obtain unique annotations (labels) for each of the considered image segments. Consequently, under such a framework, the advantages brought are threefold:

- Arbitrary image analysis algorithms can be employed for acquiring an initial set of annotations, without the need for specialised domain-tuned implementations, and integrated for achieving more complete and robust content annotations.
- The context-aware refinement of the degrees renders the annotations more reliable for subsequent retrieval steps, as the confidence is strengthened for the more plausible annotations and lowered for the less likely ones, while false annotations are reduced through the application of constraint reasoning.
- The use of ontologies, apart from allowing the sharing of domain knowledge and providing a common vocabulary for the resulting content annotations (labels),

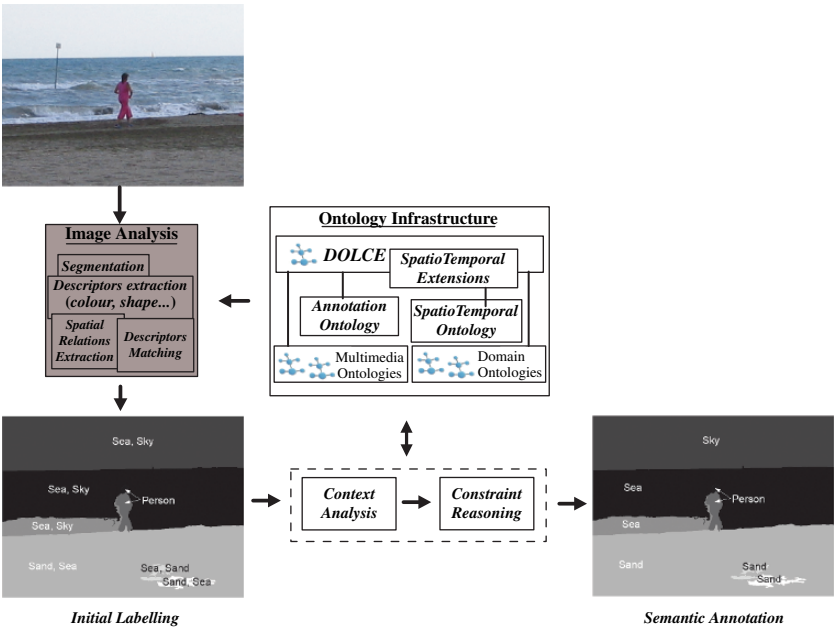


Fig. 4.1 Ontology-based framework introducing context and constraint reasoning in image analysis

ensures smooth communication among the different modules involved and facilitates interoperability with respect to future extensions with additional modules.

The rest of the chapter is organised as follows. Section 4.2 presents relevant work in terms of utilising visual context and constraint reasoning approaches in semantic image analysis, while in Section 4.3, the proposed framework is described, including the specification and design of the ontology infrastructure. Section 4.4 details the modelling and ontological representation of context of appearance and presents the methodology for readjusting the initial degrees of confidence, while Section 4.5 describes the application of constraint reasoning for the purpose of consistent image labelling. Experimental results and evaluation of the proposed framework are presented in Section 4.6, while Section 4.7 concludes the chapter.

4.2 Relevant Work

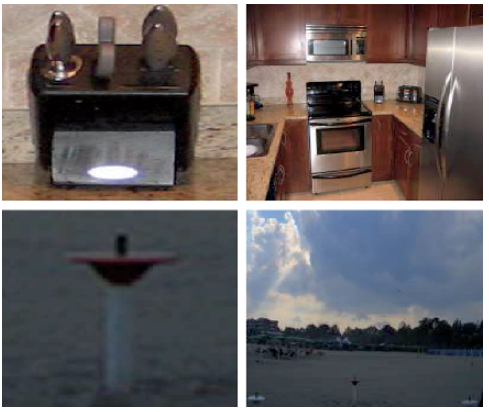
4.2.1 Context in Image Analysis

In semantic content-based image search and retrieval, research has shifted beyond low-level colour, 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 performance to be enhanced by exploiting the semantic correlation between the considered concepts. It is also rather true that in the real world, objects always exist in a context. In principle, a single image taken in an unconstrained environment is not sufficient to allow a computer algorithm or a human being to identify the object. However, a number of cues based on the statistics of our everyday visual world are useful to guide this decision. Identification of an object in an image, or a close-up image of the same object, may be difficult without being accompanied by useful contextual information. As an example, an image of a cow is more likely to be present in a landscape environment, such as a green field, whereas a desk is usually found indoors, or as depicted in Fig. 4.2, an isolated close-up picture of a kitchen gadget or beach equipment is more difficult to identify or enrol when considered out of the rest of the environmental information.

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 colour, texture, edge characteristics, and so on of the concepts considered. The advantages of context utilisation overwhelm the required effort increase on object annotation and analysis, provided a moderate balance, between the efforts spent on the identification and annotation of one object and the total amount of objects annotated within an image, will be followed. Given a particular domain, the rule of thumb, in order to obtain optimal results, is to identify a set of characteristic objects to be annotated, after statistically analysing the objects' co-occurrence in a subset of the entire dataset (e.g. 20% of the images).

A number of interesting enhanced analysis efforts have been reported including, among others, the exploitation of 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 (2003), 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



Isolated object Object in context

Fig. 4.2 Isolated object vs. object in context

information to assist in resolving local ambiguities, while in Boutell (2006), three types of context are explored for the scene classification problem, namely 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.

The aforementioned efforts indicate the shift witnessed towards utilising available contextual information in multimedia analysis. However, contrary to natural language processing (NLP), where the use of context has been investigated thoroughly (Wiebe, Hirst and Horton 1996), the respective efforts in the field of multimedia analysis are in a very early stage. The formal model of context semantics and its application as described in Section 4.4 aims to contribute with a generic methodology towards introducing and benefiting from contextual knowledge.

4.2.2 Constraint Reasoning in Image Analysis

Constraint reasoning has a long history, starting with the system SKETCHPAD (Sutherland 1963) in the early 1960s. Later, Waltz formalised the notion of constraints in order to solve the problem of deriving a 3D interpretation of 2D line drawings as the *scene labelling problem* (Waltz 1975). 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 this chapter we provide a concrete instantiation of the scene labelling problem, deployed in a real application setting. As discussed in the following, 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.

In Kolbe (1998), constraint reasoning techniques are employed for the identification of objects in aerial images. One main aspect of the presented study is the handling of over-constrained problems. An over-constrained problem is a constraint satisfaction problem in which not all constraints can be satisfied simultaneously. In traditional constraint reasoning, this would mean that no solution 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 an information theory-based evaluation measure. However, Kolbe uses, in addition, specialised constraints between the image parts that render the proposed techniques less applicable to more generic domains.

In Hotz and Neumann (2005), a configuration system is adopted to provide high-level scene interpretations. The system is evaluated on table-laying scenes, i.e. scenes where a table is laid and where the table is monitored by a camera. The goal is to identify the purpose the table is laid for, e.g. "*Dinner for Two*",

"*Breakfast*". Hotz and Neumann use well-defined domain models based on the spatial arrangements of the concepts found within the given domain to introduce reasoning into this task. 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 unlikely to exist for broad domains such as the ones of "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 "An 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, which is the case of the framework proposed in this chapter.

4.3 Ontology Infrastructure

The proposed ontology-based framework aims to serve as a generic, easy-to-extend knowledge-based framework for enhancing available semantic image analysis annotations through context-aware refinement and spatial consistency checking. As such, the intended usage purpose imposes certain requirements with respect to the knowledge infrastructure that constitutes the proposed framework's backbone, which reflect on the representation and engineering choices.

The first requirement refers to the need for smooth communication among the involved modules while preserving the intended semantics. This practically means that the annotations and the employed contextual and spatial knowledge have to be captured and represented in such a way as to promote clean semantics and facilitate exchange. The ontology languages that emerged within the Semantic Web initiative constitute promising candidates as, due to their relation with logic and particularly DLs, they provide well-defined semantics, while their XML-based syntax enhances exchange across different applications. Among the available languages, OWL DL constitutes the optimal choice with respect to expressivity and complexity trade-off. However, as described in the following sections, the expressivity requirements of the proposed framework restrict in subclass and domain/range semantics, thus not justifying the use of OWL DL or Lite. Additionally, the need for incorporating fuzziness into the representation on the one hand and the lack of a formal notation for accomplishing this on the other renders reification the only viable choice, which in turn would cancel out the inference capabilities the adoption of OWL DL would bring. For these reasons, the RDFS language was chosen for the employed knowledge infrastructure.

An additional aspect relates to the kind of knowledge that needs to be captured. Given that image analysis and annotation relate both to domain-specific aspects,

i.e. the specific domain concepts and relations, and to 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. For the multimedia-related knowledge, the MPEG-7 specifications (Sikora 2001) have been followed, as it constitutes the main standardisation effort towards a common framework for multimedia content description. Another important requirement relates to the need for enabling extensibility in terms of incorporating image analysis annotations that adhere to possibly different models of the domain or media-related knowledge. To enable the smooth harmonisation between such annotations, a reference point is needed so that the corresponding intended meanings, i.e. ontological commitments, can be disambiguated and correctly aligned. Consequently, the use of a core ontology through its rigorous axiomatisation provides the means to handle more effectively terminological and conceptual ambiguities.

As illustrated in Fig. 4.1, the developed knowledge infrastructure follows a modular architecture where different ontologies are utilised to address the different types of knowledge required. Appropriate multimedia ontologies have been developed to describe the structure and low-level features of multimedia content, which are harmonised with the corresponding domain ontologies via the use of a core ontology. The latter has been extended to cover the concrete spatiotemporal relations required when analysing such content. Finally, a dedicated ontology has been developed to provide the vocabulary and structure of the generated annotations. In the following, we briefly overview the role of each of the ontologies. For further details, the reader is referred to Bloehdorn, Petridis, Saathoff, Simou, Tzouvaras, Avrithis, Handschuh, Kompatsiaris, Staab and Strintzis (2005).

4.3.1 Core Ontology

The role of the core ontology in this framework is threefold: (i) to serve as a starting point for the engineering of the rest of the ontologies, (ii) to serve as a bridge allowing the integration of the different ontologies employed, i.e. by providing common attachment points, and (iii) to provide a reference point for comparisons among different ontological approaches. In our framework, we utilise DOLCE (Gangemi, Guarino, Masolo, Oltramari and Schneider 2002), which was explicitly designed as a core ontology. DOLCE is minimal in the sense that it includes only the most reusable and widely applicable upper-level categories, and rigorous in terms of axiomatisation, as well as extensively researched and documented.

4.3.2 SpatioTemporal 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 following Allen interval calculus (e.g. meets, before).

4.3.3 Visual Descriptor Ontology

The visual descriptor ontology (VDO) models properties that describe visual characteristics of domain objects. VDO follows the MPEG-7 visual part (ISO/IEC 2001), with some modification so as to translate the XML schema and datatype definitions into a valid RDFS representation.

4.3.4 Multimedia Structure Ontology

The multimedia structure ontology (MSO) models basic multimedia entities from the MPEG-7 MDS (ISO/IEC 2003). More specifically, the 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. Apart from the definition of classes (properties) reflecting the MPEG-7-defined descriptions, additional classes (relations) have been introduced to account for descriptions perceived semantically distinct, but treated ambiguously in MPEG-7 (such as the concept of frame).

4.3.5 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 labelling and allows the association of a degree of confidence to each label produced by the analysis.

4.3.6 Domain Ontology

In the presented multimedia annotation framework, the domain ontologies are meant to model the semantics of real-world domains that the content belongs to, such as sports events or personal holiday images. They serve a dual role: (i) they provide the vocabulary to be used in the produced annotations, thus providing the domain conceptualisation utilised during retrieval, and (ii) they provide the spatial and contextual knowledge necessary to support the context-aware and constraint reasoning refinements. As aforementioned, each domain ontology is explicitly aligned to the

DOLCE core ontology, ensuring thereby interoperability between different domain ontologies possibly used by different analysis modules.

4.4 Context Analysis

4.4.1 Ontology-Based Contextual Knowledge Representation

It should be rather clear by now that ontologies are suitable for expressing multimedia content semantics in a formal machine-processable representation that allows manual or automatic analysis and further contextual processing of the extracted semantic descriptions. Amongst all possible ways to provide an efficient knowledge representation, we propose one that relies on concepts and their relationships. In general, we may formalise domain ontologies as follows:

$$O = \{C, \{R\}\}, \text{ where } R : C \times C \rightarrow \{0, 1\} \tag{4.1}$$

where O is a domain ontology, C is a subset of the set of concepts described by the domain ontology, and R is a possible semantic relation amongst any two concepts that belong to C. In general, semantic relations describe specific kinds of links or relationships between any two concepts. In the crisp case, a semantic relation either relates ($R = 1$) or does not relate ($R = 0$) a pair of concepts 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 utilise three types of relations, whose semantics are defined in the MPEG-7 standard, namely the *specialisation* 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:

$$O_F = \{C, \{r_{pq}\}\}, \text{ where } r_{pq} = F(R) : C \times C \rightarrow [0, 1] \tag{4.2}$$

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 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), 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):

$$F = \sum_{i=1}^{n} c_i/w_i = \{c_1/w_1, c_2/w_2, \ldots, c_n/w_n\} \tag{4.3}$$

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:

$$T = Tr^t(Sp \cup P^{-1} \cup Pr^{-1}). \tag{4.4}$$

Based on the roles and semantic interpretations of Sp, P, and Pr, as they are defined in the MPEG-7 MDS (ISO/IEC 2003), it is easy to see that Equation (4.4) combines them in a straightforward and meaningful way, utilising 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, Andreou, Stamou and Kollias (2002).

The representation of this concept-centric contextual knowledge model follows the resource description framework (RDF) standard (Becket and McBride 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 visualised 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 motor-sports domain is depicted in Fig. 4.3.

Representing the graph in RDF is a straightforward task, since the 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

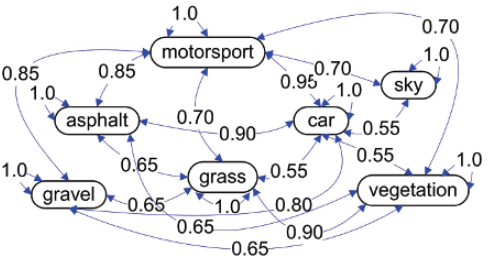


Fig. 4.3 Graph representation example – motor-sports domain

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 *Motor-sportsScene part Car*, which means that a car is part of a motor-sports scene, and a membership degree of 0.75 for this statement does obviously not entail that a car is always a part of a motor-sports scene. A small illustrative example is provided in Table 4.1 for an instance of the specialisation 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 specialisation are subclassing, i.e. x is a generalisation of y, and thematic categorisation, i.e. x is the thematic category of y. In the example, the RDF subject *wrc* (World Rally Championship) has *specialisationOf* as an RDF predicate and *rally* forms the RDF object. Additionally, the proposed reification process introduces a statement about the former statement on the *specialisationOf* resource, by stating that 0.90 is the membership degree to this relation.

4.4.2 Visual Context Analysis

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, as the latter is defined on the "fuzzified" version of traditional ontologies presented in Section 4.4.1. From a practical point of view, we consider context as

Table 4.1 Fuzzy relation representation: RDF reification

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

information depicted by specific domain concepts that are identified and whose relations are analysed based on the utilised data set and not by external factors, such as EXIF metadata.

In a more formal manner, the problem that this work attempts to address is summarised 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. In designing such an algorithm, contextual information residing in the aforementioned domain ontology is utilised. In general, the notion of context is strongly related to the notion of ontologies since an ontology can be seen as an attempt towards modelling real-world (fuzzy) entities, and context determines the intended meaning of each concept, i.e. a concept used in different contexts 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 the previous sections, 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 with a region a of an image in an image scene. Each specific concept $k \in C$ present in the application domain's ontology is stored together with its relationship degrees r_{kl} to any other related concept $l \in C$.

Another important point to consider is the fact that each concept has a different probability of appearing 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. 4.3, every concept participating in the contextualised 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 motor-sports element could be related either directly or indirectly with any other concept. To tackle cases where 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 characterising each domain dm. For instance, the root element of the motor-sports domain is concept $c_{motorsports}$. 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 degree of membership 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 motor-sports domain ontology, assuming that the

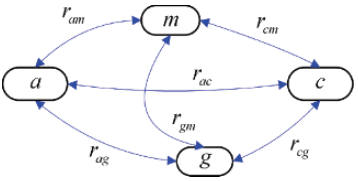


Fig. 4.4 Graph representation example – compatibility indicator estimation

only available concepts are *motorsports* (the root element – denoted as m), *asphalt* (a), *grass* (g), and *car* (c), is presented in Fig. 4.4 and summarised 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 $cr_{dm}(a)$:

$$cr_{dm}(a) = \max\{r_{am}, r_{ag}r_{gm}, r_{ac}r_{cm}, r_{ag}r_{cg}r_{cm}, r_{ac}r_{cg}r_{gm}\}. \tag{4.5}$$

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

1. Identify an optimal normalisation 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 with 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 with the region, based on np and the context's relevance value. In the case of multiple concept relations in the ontology, when relating concept k to more than one concept, 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:

$$\mu_a^n(k) = \mu_a^{n-1}(k) - np(\mu_a^{n-1}(k) - cr_k) \tag{4.6}$$

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

$$\mu_a^n(k) = (1 - np)^n \cdot \mu_a^0(k) + (1 - (1 - np)^n) \cdot cr_k \tag{4.7}$$

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

In practice, typical values for n reside between 3 and 5. Interpretation of the above equations implies that the proposed contextual approach will favour confident degrees of membership for a region's concept in contradistinction 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. Furthermore, based on the supplied ontological knowledge, it will clarify and solve ambiguities in cases of similar concepts or difficult-to-analyse regions.

A key point in this approach remains the definition of a meaningful normalisation 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 an 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, utilising semantics in the form of the contextual information that resides in the constructed "fuzzified" ontology.

4.5 Constraint Reasoning to Eliminate Ambiguities in Labelled Images

So far, the initial labelling provides a hypothesis set of labels for each segment, that is computed based on the low-level features extracted from the specific segment. Each label is associated with a degree of confidence, indicating how likely the label is to be depicted. The context algorithm introduces global context into the labelling by readjusting the degrees for each label. In this section, we will discuss the application of spatial knowledge to the initially labelled image, with the goal to identify a final and spatially consistent labelling. The spatial knowledge will be represented by a set of spatial constraints, and the initially labelled image will be transformed into a *constraint satisfaction problem (CSP)*, which will be solved using standard constraint reasoning techniques.

4.5.1 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. 4.5, 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

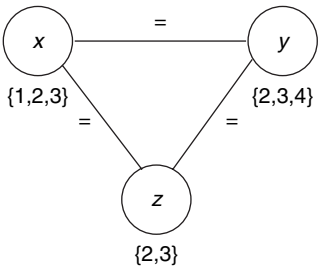


Fig. 4.5 A simple constraint satisfaction problem

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 $c_j = D(v_1) \times \cdots \times D(v_r)$ 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 problems, 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 its constraints is arc consistent.

Now, in the example of Fig. 4.5, 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 has support neither 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 a 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, which prune the search space during search using local consistency notions and thus provide an improved runtime performance. 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 this chapter. 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).

4.5.2 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.

Spatial relations provide an important means to interpret images and 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 an initially labelled image into a CSP.

4.5.2.1 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. Furthermore, 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_V^t of type t on the set of variables V is now defined as $c_V^t := D(t) \cap (D(v_1) \times \cdots \times D(v_n))$. Apparently, c_V^t now is a relation on the variable domains containing only those tuples that are allowed for the spatial relation r_t.

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, such as *contained-in* or *above-of*. Absolute spatial constraints are derived from the absolute positions of segments on the image, such as *above-all*, and which are apparently unary constraints.

4.5.2.2 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 straightforward process. For each segment, a variable is created and the hypotheses sets become the domains of the variables. For each spatial relation, a constraint with the corresponding type is added. In the following, we will formalise the transformation.

Let $I = (S, SR)$ be a 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. For the newly created variable v^s, set the domain to $D(v^s) = ls(s)$.
3. Let SR be the set of all spatial relations defined in the domain knowledge, then 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 4.5.1. 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, 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.

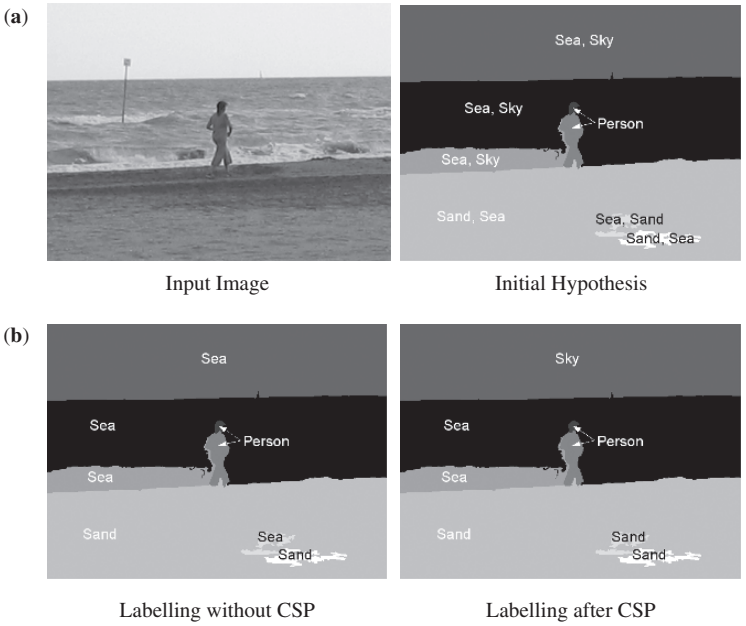


Fig. 4.6 Example of CSP application

An example is depicted in Fig. 4.6, where the input image, the initial set of hypotheses, the corresponding labelling that would have been produced without constraint reasoning, and the labelling after the constraint reasoning are depicted. Please 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 region 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.

4.6 Experimental Results and Evaluation

In this section, we present experimental results and evaluation of the enhancement achieved by the application of the proposed context analysis and constraint reasoning modules over typical image analysis. As aforementioned, under the proposed framework, image analysis is treated as a black box, and different implementations can be used interchangeably. In the presented experimentation, we followed the approach presented in Petridis, Bloehdorn, Saathoff, Simou, Dasiopoulou, Tzouvaras, Handschuh, Avrithis, Kompatsiaris and Staab (2006) for two main reasons:

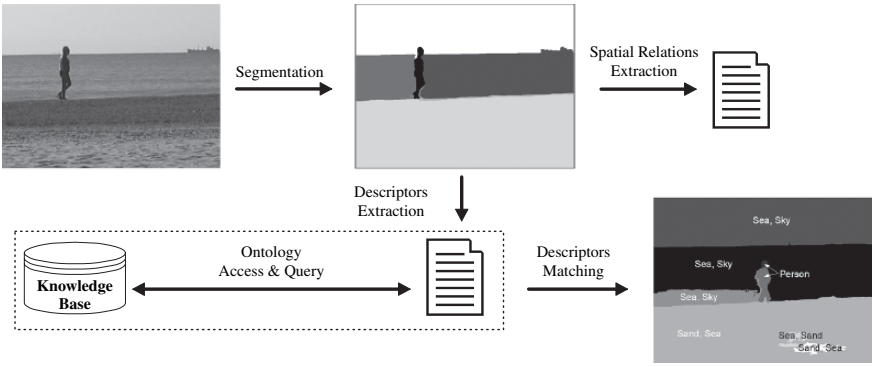


Fig. 4.7 Image analysis architecture

(i) the presented approach is quite generic, not making use of domain-specific implementation and tuning that would boost performance, and (ii) the produced annotations adhere to the proposed framework ontology infrastructure, thus making the application of the framework straightforward, without the need for an intermediate aligning step to harmonise the annotations' semantics with the corresponding framework.

The overall architecture of the analysis used for experimentation is illustrated in Fig. 4.7. First segmentation is applied to partition the image into a set of segments. Subsequently, for each of the resulting segments, the dominant colour, homogeneous texture, and region shape descriptors are extracted, and additionally, the spatial relations between adjacent segments are estimated. Initial sets of graded hypotheses, i.e. sets of labels with associated degrees of confidence, are generated for each image segment through the computation of matching distances between each segment's descriptors and the prototypical values defined for the considered domain objects. These prototypical values are created using the M-OntoMat-Annotizer tool, which enables users to annotate segments with concepts from a given ontology and then extract selected descriptors, thus allowing the linking of low-level visual features to domain concepts (Petridis, Bloehdorn, Saathoff, Simou, Dasiopoulou, Tzouvaras, Handschuh, Avrithis, Kompatsiaris and Staab 2006).

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, as well as to statistically induce the initial fuzzy values of relations utilised within the context ontology. The resulting 120 images have first undergone the aforementioned analysis in order to obtain the corresponding initial annotation (labelling). Then, the proposed framework was applied. First, the context analysis module, exploiting the domain concepts' associations and the information extracted through training, readjusts the annotations' degrees of confidence towards more meaningful values. Secondly, the constraint reasoner, applying the spatial rules on the contextually refined labels results in the removal of those that violate the domain spatial topology. To quantify the performance of image analysis,

and allow us to measure the enhancement brought by the proposed framework, we keep for each image segment the label with the highest degree of confidence from the respective hypotheses set as the analysis results. Similarly, to measure the performance of context analysis and constraint reasoning, the label with the highest degree is kept for each segment.

To overcome the difficulties and cost in defining generally accepted pre-annotated segmentation masks and avoid getting into a 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 a certain tolerance for these kinds 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 smaller the block size, the greater 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.

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 formulae, 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 the arithmetic mean, it gets large only if both precision and recall are large. In the case that 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. Furthermore, objects that appear in the test images but do not belong to the supported set of concepts have not been taken into account, since they do not add to assessing the proposed modules performance.

In the current experimentation, six concepts have been considered, namely *Cliff*, *Person*, *Plant*, *Sand*, *Sea*, and *Sky*. In Table 4.2, the precision (p), recall (r), and *F-measure* (f) are given for the examined test images with respect to sole image analysis, image analysis followed by context analysis, and image analysis followed by constraint reasoning respectively, while in Table 4.3, the integrated performance is shown. From the obtained results, one easily notes that in almost 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, a lower improvement is observed with respect to the concept *Person*, as due to over- and under-segmentation phenomena the effects of the transition from 2D to 3D, and its generic context of appearance a region depicting a *Person* may validly appear almost in any configuration with respect to the rest of the domain concepts. Observing the integrated context

Table 4.2 Evaluation results for the beach domain (where IA, CTX and CSP stand for image analysis, context, and constraint reasoning, respectively)

	IA			IA+CTX			IA+CSP		
Concept	*p*	*r*	*f*	*p*	*r*	*f*	*p*	*r*	*f*
Cliff	0.09	0.20	0.12	0.30	0.94	0.46	0.47	0.40	0.44
Person	0.56	0.40	0.47	1.00	0.07	0.14	0.61	0.40	0.48
Plant	0.35	0.77	0.48	0.72	0.26	0.38	0.85	0.89	0.87
Sand	0.82	0.80	0.81	0.90	0.95	0.92	0.81	0.94	0.87
Sea	0.87	0.58	0.70	0.90	0.83	0.86	0.87	0.49	0.63
Sky	0.86	0.89	0.87	0.94	0.94	0.94	0.80	0.95	0.87
AVG	0.73	0.73	0.73	0.84	0.85	0.84	0.77	0.75	0.76

Table 4.3 Evaluation results for the beach domain for the combined application of context and constraint reasoning over image analysis

	IA+CTX+CSP		
Concept	*p*	*r*	*f*
Cliff	0.38	0.94	0.54
Person	1.00	0.14	0.25
Plant	0.82	0.48	0.61
Sand	0.90	0.97	0.93
Sea	0.90	0.86	0.88
Sky	0.95	0.91	0.93
AVG	0.86	0.86	0.86

analysis and constraint reasoning results, it is noted that the latter adds only a little to the attained performance, compared to when combined with image analysis only. However, given the set of concepts currently supported and the inaccuracies of the segmentation, this is an expected outcome. Having a broader set of concepts from different and possibly partial overlapping domains (in terms of concepts included) would lower the context refinement accuracy and would make more evident the role of spatial consistency for disambiguation.

4.7 Conclusions and Further Discussions

In this chapter, we have proposed an ontology-based framework for enhancing semantic image analysis through the refinement of initially available annotations by means of explicit knowledge about context of appearance and spatial constraints of the considered semantic objects. 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 can be summarised as follows: (i) the formal representation of context of appearance semantics in an ontology compliant way that facilitates its

integration within knowledge-based multimedia analysis, and a methodology for its application; (ii) the adoption of a constraint problem solving methodology within the semantic image annotation domain for addressing topological knowledge; and (iii) the proposed framework that supports its applicability and extensibility to different image analysis applications.

Future directions include further investigation of the proposed framework using more concepts, thereby making available additional knowledge, i.e. more spatial constraints and contextual associations. More specifically, with respect to the constraint reasoner, a fuzzified extension is under investigation in order to provide greater flexibility and better scalability to broader domains. Introducing such uncertainty support will enable the handling of situations that cannot be adequately modelled in the provided domain knowledge, and for which the current crisp implementation may fail to provide a solution, i.e. none of the values may satisfy the constraints. Another appealing characteristic of using a fuzzy CSP approach is that preferences among certain solutions can be captured, as for instance solutions where the sea is above the sand. Furthermore, since the manual definition of constraints for large numbers of concepts is infeasible and error-prone, a heuristic approach towards a more efficient acquisition needs to be investigated. With respect to contextual knowledge modelling and utilisation, an interesting future aspect refers to the exploration of additional semantic associations between the concepts that participate in a domain and the interdependencies that emerge from overlapping sets of concepts between different domains. Finally, experimentation with alternative analysis modules or their combination would provide useful and concrete insight into the proposed framework contribution in real applications scenarios.

Acknowledgments The work presented in this chapter was partially supported by the European Commission under contract FP6-001765 aceMedia.

References

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

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

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

Becket, D. and McBride, B. (2004) *RDF/XML Syntax Specification, W3C Recommendation, 10 February*.

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: Proceedings of the 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.

Boutell, M. and Luo, J. (2005) *Beyond Pixels: Exploiting Camera Metadata for Photo Classification*. In: Pattern Recognition 38(6).

Brickley, D. and Guha, R.V. (2004) *RDF Schema Specification 1.0, W3C Recommendation, 10 February*.

Dasiopoulou, S., Mezaris, V., Kompatsiaris, I., Papastathis, V.K., and Strintzis, M.G. (2005) *Knowledge-Assisted Semantic Video Object Detection*. In: IEEE Transactions 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.

ISO/IEC (2001) 15938-3:2001: *Information Technology – Multimedia Content Description Interface – Part 3 visual.* Version 1.

ISO/IEC (2003) 15938-5:2003: *Information Technology – Multimedia Content Description Interface – Part 5: Multimedia Description Schemes.* First Edition.

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. (2003) *Natural Object Detection in Outdoor Scenes Based on Probabilistic Spatial Context Models*. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), pp. 457–461.

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.

Mylonas, Ph. and Avrithis, Y. (2005) *Context modeling for multimedia analysis and use*. In: Proceedings 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: Advances in Neural Information Processing Systems 16 (NIPS), Vancouver, BC, MIT Press.

Naphade, M., Kozintsev, I. and Huang, T.S. (2002) *Factor Graph Framework for Semantic Indexing and Retrieval in Video*. In: IEEE Transactions on Circuits Systems Video Technology, 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.

Petridis, K., Anastasopoulos, D., Saathoff, C., Timmermann, N., Kompatsiaris, I., and Staab, S. (2006) *M-OntoMat-Annotizer: Image Annotation. Linking Ontologies and Multimedia Low-Level Features*. In: 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES 2006), Bournemouth, UK, October.

Petridis, K. Bloehdorn, S., Saathoff, C., Simou, N., Dasiopoulou, S., Tzouvaras, V., Handschuh, S., Avrithis, Y., Kompatsiaris, I., and Staab, S. (2006) *Knowledge Representation and Semantic*

Annotation of Multimedia Content. IEE Proceedings on Vision Image and Signal Processing, Special issue on Knowledge-Based Digital Media Processing, Vol. 153, No. 3, pp. 255–262, June.

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 Transactions 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: PhD thesis, Massachusetts Institute of Technology.

Vailaya, A. and Jain, A. (2000) *Detecting Sky and Vegetation in Outdoor Images*. In: Proceedings of 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.

Wiebe, J., Hirst, G., and Horton, D. (1996) *Language Use in Context*. In: Communications of the ACM, 39(1), pp. 102–111.