Semantic Annotation for Retrieval of Visual Resources

Laura Hollink

SIKS Dissertation Series No. 2006-24

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Promotiecommissie:
prof.dr. A.Th. Schreiber (promotor)
prof.dr. B.J. Wielinga (promotor)
dr. M. Worring (copromotor)
dr. L.M. Aroyo
prof.dr. P.G.B. Enser
prof.dr. F.A.H van Harmelen
prof.dr. E. Hyvönen
prof.dr.ir. A.W.M Smeulders

ISBN 90-8659042-X

VRIJE UNIVERSITEIT

Semantic Annotation for Retrieval of Visual Resources

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Exacte Wetenschappen
op donderdag 16 november 2006 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door
Laura Hollink
geboren te Haarlem

promotoren:	prof.dr. A.Th. Schreiber
	prof.dr. B.J. Wielinga
copromotor:	dr. M. Worring

Contents

1 Introduction **1**

1.1 Context 1
1.2 Research Questions 4
1.3 Approach 5
1.4 Overview of the Thesis 6

2 Classification of User Image Descriptions **9**

2.1 Introduction 9
2.2 Related work 10
2.3 A Framework for the Classification of Image Descriptions 12
2.4 User Study 21
2.5 Results 26
2.6 Discussion 30

3 Evaluating the Application of Semantic Inferencing Rules to Image Annotation **33**

3.1 Introduction 33
3.2 Related Work 34
3.3 The Domain of Pancreas Cells 35
3.4 Vocabularies 37
3.5 Semantic Inferencing and the Rules-By-Example Interface 39
3.6 Previous Work 39
3.7 Evaluation 42
3.8 Discussion 45
3.9 Future Work 46

4 Assessing User Behaviour in News Video Retrieval **49**

4.1 Introduction 49
4.2 The Interactive Video Retrieval System 50
4.3 Related Work: A Comparison to Other Systems 53
4.4 Methods 55
4.5 Subjects 56
4.6 Results 58
4.7 Discussion 63

5 Semantic Annotation of Image Collections **65**
5.1 Introduction . . . 65
5.2 Related Work . . . 67
5.3 Ontologies and Metadata Schema . . . 68
5.4 Links Between Ontologies . . . 72
5.5 Annotation and Search Scenario's . . . 74
5.6 Discussion . . . 79

6 Query Expansion for Image Content Search **81**
6.1 Introduction . . . 81
6.2 Related Work . . . 82
6.3 Annotation with the E-Culture Web Demonstrator . . . 84
6.4 Experimental Setup . . . 89
6.5 Results . . . 90
6.6 Discussion and Conclusion . . . 94

7 Adding Spatial Semantics to Image Annotations **99**
7.1 Introduction . . . 99
7.2 Related Work . . . 100
7.3 Representation of Spatial Concepts . . . 101
7.4 Spatial Annotation Tool . . . 102
7.5 Preliminary Evaluation . . . 104
7.6 Discussion . . . 107

8 Building a Visual Ontology for Video Retrieval **109**
8.1 Introduction . . . 109
8.2 Related Work . . . 110
8.3 Requirements . . . 111
8.4 Design of a Visual Ontology . . . 112
8.5 A Thought Experiment . . . 115
8.6 Discussion . . . 117

9 Conclusions and Discussion **119**
9.1 Research Questions Revisited . . . 119
9.2 Discussion and further research . . . 125

A Guidelines for the Classification of Image Descriptions **127**

B E-Culture Annotation Template **131**

Bibliography **137**

Samenvatting **149**

SIKS Dissertation Series **153**

Preface

At first glance, a PhD thesis looks like one coherent piece of work. Further examination, however, reveals that many theses are a collection of a diverse range of work: work done in the author's naive first year, later work, work done with a variety of people, at a number of locations, on various domains. Combining these pieces of work into a thesis is not a well-structured process. Rather, it is like solving a jigsaw puzzle that has round, square, donut shaped, and pear shaped pieces. This final task of putting together the pieces of the puzzle was the most challenging phase of my PhD, but also the one I enjoyed most. Now it is time to thank the people that have helped me with one or more pieces of the puzzle.

Firstly, I would like to thank my supervisors Guus Schreiber, Bob Wielinga and Marcel Worring. Thanks to Guus for all his time, support and feedback. Guus has the ability to transform a vaguely formulated problem into a clear and solvable question, and I always left our meetings feeling more optimistic then when I entered them. I am grateful to Bob for giving me the freedom to move to the VU. Bob's critical reading has been essential to finishing this thesis, especially in the last months. I'd like to thank Marcel, who has the rare gift of explaining his work to people outside his field in an understandable way. This helped make our collaboration to combine our two separate fields an interesting one. Also thanks to Giang Nguyen for her cooperation and for her pleasant company on SIKS courses and our presentation course.

I very much enjoyed the friendly atmosphere at the former SWI department of the UvA. Thanks to Noor for making me feel welcome when I first started as a PhD student. After working at the UvA for more than a year I moved to the VU. There, the BI and KR groups provided a lively working environment with lots of presentations and discussions, but also movie nights, the 'klimlijst' and sailing trips. Special thanks to Mark, who has become a real friend and, because of his precise way of working, also a valuable colleague. In the winter of 2005 I spent a lovely summer in Brisbane, Australia with Jane Hunter and her group. Thanks to Jane and Suzanne for the pleasant and productive cooperation. Also thanks to Ronny, Sarah and Phil for the good company.

I wish to thank Anna and Lennart. Who would have thought that after all these years we would come to understand each other's work? I can't wait for their theses to be ready. Many thanks to Joanne for reading and correcting parts of this thesis.

I would like to thank Karel and Wil for their continuous interest in the progress of my work, which has been a great comfort. I was motivated by their systematic efforts to store, digitise and eventually even index all our holiday pictures! Thanks to Vera for proof-reading and help with typing, and for being a fellow PhD student with whom I could talk, gossip, and complain about the day-to-day troubles of the job. I promised Alistair not to thank him for his help with latex, statistics, the front cover, etc., so I won't. Alistair, thank you for showing me that many things, and certainly thesis writing, are easier and better when you are truly together with someone.

1

Introduction

1.1 Context

Images are used for a variety of purposes: designers use them as decoration, editors as illustration, teachers use for explanation, students as sources of information and scientists use them to prove a point. Locating an image that meets a user's need in a large and varied collection is a difficult and time consuming task. Many have noted the ever increasing size of image and video collections (e.g. Jaimes and Chang 2000, Jörgensen 1999). This growth increases the size of the proverbial haystack of images, through which a user has to search. As a result, images become harder to find. The problems that users face when searching for an image and the need for efficient retrieval methods are becoming widely recognised (e.g. Mezaris et al. 2004, Smeulders et al. 2000). The retrieval of still and moving images, or *visual resources*, is the central topic of this thesis.

One way to enable retrieval of visual resources from large collections, is by annotation. An annotation has been loosely defined by the World Wide Web Consortium (W3C) Annotation Working Group (1995) as "any object that is associated with another object by some relationship". In the Annotea project, a more specific definition of an annotation was given: "comments, notes, explanations or other types of external remarks that can be attached to any web document or a selected part of the document without actually needing to touch the document" (Koivunen 2005). According to the Oxford Advanced Learners Dictionary, annotation means: "to add notes to a book or text giving explanations or comments". These definitions can easily be extended to include visual documents and to allow other types of information than just explanations, comments and notes. However, none of these definitions include the reason for annotation. Therefore, in this thesis, we define an annotation as *information that is explicitly related to an item with the purpose of describing the item for future reference and retrieval*.

The question of what information should be captured in an annotation of a visual resource is a non-trivial one, since the meaning of a visual resource depends for a large part on the interpretation of the person looking at it. One visual resource can be described in different ways by different people, depending on the person's goal, background and expertise. A photo of the fall of the Berlin Wall, for example, can be described as "a black-and-white photo created by A. N. de Wit in November 1989", "people, night, concrete", "a man climbing a wall" or "the Iron Curtain". An annotation of a visual resource should convey the broad range of queries that people might formulate to retrieve that resource.

Annotation practices range from unstructured annotation, in which, for example, free text is used to describe images, to highly structured annotation with controlled vocabularies and metadata schemas. A popular example of unstructured annotation is Flickr (http://www.flickr.com/). Flickr is a web application in which members can upload photos and annotate them with 'tags'. There are no restrictions on the form and the amount of tags associated to a photo; the tags can be anything, even non-existent words. If the owner of a photo allows it, every member of Flickr can add tags to the photo. The purpose of Flickr is to facilitate photo management and sharing more than to facilitate search. For sharing purposes, the lack of structure is a feature rather than a bug: it makes the site easy to use and gives users freedom of expression in any language. However, search through the entire collection is difficult because of the inconsistency in tag use. Photos of the International Semantic Web Conference in 2005, for example, are tagged with "iswc AND 2005" (2 photos), "iswc2005" (240 photo's) and "iswc05" (35 photos), making the selection of an adequate search term difficult.

Controlled vocabularies can introduce coherency in the use of annotation and query terms. Controlled vocabularies can be simple lists of words (e.g. Volkmer et al. 2005) or more structured resources such as lexical databases, thesauri or ontologies. The most widely-known lexical database is WordNet. It is a "semantic dictionary" (Fellbaum 1998, p. 7) that organises words based on their meaning and use in natural language. A thesaurus is a vocabulary of terms with hierarchical, associative and equivalence relations between them (Assem, van et al. 2004). The Getty Institute, for example, publishes three thesauri, containing "terms, names and other information about people, places, things and concepts relating to art, architecture and material culture", that can be used as annotation and search vocabularies (The Getty Foundation, 2006b). An ontology is defined by Gruber (1993, p. 1) as "an explicit specification of a conceptualisation", where a conceptualisation consists of "the objects, concepts and other entities that are presumed to exist in some area of interest and the relationships that hold them". In other words, "a conceptualisation is an abstract, simplified view of the world that we wish to represent for some purpose". The relations in an ontology are more formally defined than relations in a thesaurus.

Ontologies, thesauri and lexical databases are not clearly distinguishable from each other as pointed out by Smith and Welty (2001). Rather, their instances can be ordered along a continuum of increasingly formally-structured vocabularies. Throughout the chapters of this thesis, we use the terms 'thesaurus' and 'ontology' to denote annotation- and search-vocabularies.

Annotation using concepts from existing, well established vocabularies has advantages over annotation using an in-house vocabulary. The semantics of vocabularies such as the Getty thesauri and WordNet are widely agreed upon. If annotation concepts are taken from these vocabularies, others can use the semantic structure from the underlying vocabulary to interpret the annotation concepts.

Metadata schemas are a way to further structure the use and interpretation of a vocabulary. They consist of elements or properties that indicate the way terms in the vocabulary are linked to the visual resource. With a metadata schema one can, for example, distinguish a painting currently being displayed in Paris from a painting depicting Paris. A widely known example of a metadata

schema is the Dublin Core metadata standard (Dublin Core 2006). Dublin Core (DC) provides 15 elements to describe resources, such as creator, date and type. Similar to the Dublin Core elements, but focussing only on visual resources, are the VRA Core Categories. They are an ISO standard for the description of various aspects of visual resources, such as the creator, date, title and measurements (VRA 2002). They provide a mapping to Dublin Core elements so that the elements can always be 'dumbed down' (i.e. reduced) to Dublin Core elements. Some specific meaning will be lost but the value of the element is still generally correct. The use of standard metadata schemas increases the interoperability of annotations. A domain in which this highly structured type of annotation is frequently used, is the cultural heritage domain (Graham 1999). Structured annotation with controlled vocabularies and metadata schemas has been, and still is, a way for museums to store information about visual resources in their collection and as a result ensure access to these resources.

Advances in knowledge representation, mainly in the area of the semantic web, make new ways of sharing annotations possible. Standardised languages, such as the Resource Description Framework (Schema) (RDF(S)) (Brickley and Guha 2000) and the Web Ontology Language OWL (Web Ontology Working Group 2003), have been developed to formally describe concepts and relations and properties thereof. The constructs in these languages have predefined semantics and resources described in these languages are uniquely identified by Uniform Resource Identifiers (URI's). One can express, for example, properties of resources, restrictions on the value of these properties, disjointness statements and logical relationships between resources (Antoniou and Harmelen, van 2004). The formal semantics of RDF/OWL make it possible for machines as well as humans to share, process and reason with the information described in these languages.

Many of the metadata standards and ontologies that have been developed for annotation and search are now being translated into the aforementioned semantic web languages. Concepts in the ontologies become uniquely identified, while relations between concepts get explicit meaning. This opens up the possibility to use semantic web techniques for annotation and search. Annotation concepts become processable for annotation- and search-tools. With semantic web techniques, the image of the Berlin Wall will still be described in different ways by different people. However, if these descriptions are made up of concepts from a well-known ontology in RDF/OWL, people will be able to understand, interpret and even reason with each others descriptions.

Annotation is a time consuming process. This has given rise to the question whether part of the annotation process can be automated. Image analysis is used to automatically create annotations based on the intrinsic, low-level properties of the image itself, such as colours, shapes and textures. Large training collections are used to associate combinations of image properties with annotation concepts. This approach is efficient since one can query for images that have not been manually annotated. Another method to automatically retrieve images is query-by-example. In this approach, images are retrieved based on their low-level visual similarity to an example image. Retrieval that is based on low-level visual features of an image is called content-based image retrieval (CBIR). CBIR techniques suffer from the *semantic gap*, which is the discrepancy between the information that can be derived from the low-level image data and the interpretation that users

have of an image. The description “Iron Curtain”, for example, can never be derived from the visual properties of the image alone and also a more general description like “a man climbing a wall” is not feasible with current CBIR techniques. Due to the semantic gap, the creation of high-level annotations for a broad domain with fully automatic CBIR methods is not in sight.

Despite the development of controlled vocabularies in semantic web languages and advances in the field of CBIR, the problems that users face when searching through large, heterogenous image collections remain considerable. There is a lack of synergy between the various approaches to attack the image retrieval problems. This has increased our belief that an interdisciplinary approach is needed. Although the focus of the thesis is on the use of structured background knowledge, solutions are sought not only in this area, but also in the combination with CBIR.

1.2 Research Questions

The process of annotating and searching for visual-resources remains difficult and time consuming for annotators and searchers. In this thesis, we address this problem. We take into account the whole retrieval process, including user information needs, query formulation, annotation and search. The research in this thesis aims to contribute to the general research question:

How can the process of visual-resource annotation and search be enhanced?

We refine this question in four specific research questions. Before investigating possible solutions we heighten our understanding of the process by studying two typical problems of visual-resource retrieval. Firstly, the variety of interpretations that are possible of one image make annotation difficult. If an image has been annotated based on one interpretation but a query is formulated based on another interpretation, the image will not be found. Knowledge of the types of descriptions that can be formulated for a visual resource and knowledge of the types of descriptions that are formulated in practice, are a prerequisite for solutions that support manual annotation as well as automatic annotation. Therefore, the first ‘problem oriented’ question is:

1 *How do people describe and search for visual information?*

Secondly, automatic annotation is limited by the semantic gap. The extent of this limitation, however, is not clear. To enhance the retrieval process for the user, it is necessary to know the possibilities of automatic annotation, despite the semantic gap. In order to predict how big the influence of the semantic gap is in different domains, we ask the second problem oriented question:

2 *What are the circumstances under which the semantic gap limits retrieval?*

The two problem oriented questions provide requirements and input for two subsequent ‘solution oriented’ questions that each address a different approach to enhance the process of visual-resource retrieval for the user. People use their background knowledge about the content of an image in their interpretation of that image. To a certain extent, background knowledge about a domain is

available in the ontologies that have been used as controlled vocabularies. Putting this structured background knowledge to use for retrieval could help to interpret annotation and query terms. This leads to the following solution oriented research question:

3 *How can structured background knowledge about the domain be used to support the process of visual-information retrieval?*

An interdisciplinary approach that combines structured background knowledge and image analysis techniques might further enhance the process by lightening the burden on the annotator. The fourth research question is:

4 *How can structured background knowledge and image analysis techniques be combined to improve visual-information retrieval?*

1.3 Approach

The research questions above are studied in three domains of visual information: organic cells (Chapter 3), paintings (Chapters 5, 6 and 7), and broadcast news (Chapters 4 and 8). The first is a domain with a relatively clear link between low-level visual features and high-level concepts, which makes it a good starting point to explore the limitations caused by the semantic gap. The second is a knowledge-rich domain, in which large bodies of structured background knowledge are available and experts agree on the main concepts and relations. This makes it possible to investigate the benefits of applying background knowledge to the retrieval process. The last is a broad and complex domain, in which the semantic gap complicates retrieval. The painting and cell domains both consist of still images, while the news domain contains video. However, the temporal aspect of the news videos was only marginally used. To a large extent, key frames (which are still images) of the shots were used instead of the sequence of frames that make up the video. Therefore, in this thesis we do not discuss the difference between still and moving images. The specific characteristics of each domain and the generalisation of the results to other domains are treated in the corresponding chapters.

The first question regarding descriptions of visual resources is addressed by building a framework in which a broad range of descriptions can be classified. The framework is based on a literature study. To determine which (combinations of) classes of the framework are frequently used in practice, descriptions of visual resources are studied in three contexts: in a setting not related to a specific domain or retrieval method, in the domain of paintings and in the domain of broadcast news using a CBIR system.

Question two about the nature of the semantic gap is first treated by investigating direct links between low-level visual features and high-level concepts in the domain of organic cells. It is determined what the characteristics of a domain are that make this direct link possible. With these characteristics in mind, we explore the possibilities of a direct link in the painting domain: we develop and evaluate a semi-automatic annotation system for a subset of painting annotations,

namely spatial descriptions of objects. In the complex domain of broadcast news, the direct link is difficult to make; it is known that the semantic gap plays a big role here. A user study was undertaken to find out whether characteristics of queries (amongst other factors) determine the size of the semantic gap.

Question three studies how structured background knowledge can aid retrieval. The structure, content and format of available background knowledge in the paintings domain is analysed and put to use in an annotation and search system. By means of 'use cases' the benefit of background knowledge for speeding up annotation and improving search results is shown. In an experiment, the benefit of background knowledge for content search is quantified in terms of increase in recall and precision.

The last question regarding the combination of background knowledge and image analysis techniques is addressed in three steps. Firstly, we explored how a user can make this combination. We present and evaluate an approach in which users formulate semantic web rules to translate low-level visual characteristics of images into high-level annotations. Secondly, we studied how background knowledge and image analysis techniques can be combined using an ontology. Knowledge about the visual properties of concepts in the news domain are incorporated in a high-level ontology. The performance of this 'visual ontology' in a retrieval task is evaluated in a thought experiment. Thirdly, the combination is made in the domain of paintings by automatically extending manual annotations with extra information.

1.4 Overview of the Thesis

This thesis is divided into nine chapters. Roughly, Chapters 2, 3 and 4 address questions 1 and 2, which are the two problem oriented questions. Chapters 5 and 6 address the third question and Chapters 7 and 8 address question 4. Finally, Chapter 9 provides the conclusions. The chapters are self contained and can be read independently from each other.

In **Chapter 2**, a framework is developed for the classification of image descriptions by users, based on various classification methods in literature. The framework was used in an empirical study to determine how the classes of the framework are used in practice.

Chapter 3 describes a method for semi-automatic annotation of images and evaluates it on images of organic cells. The performance of this approach in the pancreatic cell domain is compared to previous results in the less complex fuel cell domain. From this comparison, characteristics of a domain are derived that indicate whether the method will or will not work in a domain.

Chapter 4 presents the results of a user study, in which subjects queried a news archive using an interactive video retrieval system. Questionnaire data, logged user actions on the system, queries formulated by users and a quality measure of each search were studied. Based on the results, implications for the design of user interfaces of video retrieval systems are discussed.

In **Chapter 5**, a tool for semantic annotation and search in a collection of art images is discussed. Multiple existing ontologies are used to support this process. We discuss knowledge-engineering aspects such as the annotation structure and links between the ontologies. The frame-

work for classification of image descriptions (Chapter 2) serves as input to the annotation template of the tool. The annotation and search process is illustrated with application scenarios.

Chapter 6 reports on an experiment in which the benefit of using an ontology for content search is measured. The annotations that were used in the experiment were made on the E-Culture web demonstrator, a system that builds on the experiences of the research described in Chapter 5. Annotation properties of the demonstator are discussed.

Chapter 7 discusses support of users in adding spatial information semi-automatically to annotations of images. Existing semantic annotations of objects depicted in an image are extended with information about the absolute and relative positions of those objects. We report on a small evaluation study in which annotations generated by the tool are compared to manual annotations by ten subjects.

In **Chapter 8**, the requirements of a visual ontology are identified. Based on these requirements, a visual ontology is created out of two existing ontologies (WordNet and Mpeg-7) by creating links between visual and general concepts. Performance of the visual ontology is tested on 40 shots of news video, and the added value of each visual property in the ontology is discussed.

Finally, **Chapter 9** summarises the main conclusions.

2

Classification of User Image Descriptions

One visual resource can be described in many different ways. In this chapter we study the different categories of image descriptions. We report on a literature study regarding image classification schemes, methods and standards. A framework for the classification of images is constructed that integrates the various viewpoints found in literature. In an empirical study, we classify image descriptions of 30 participants using the framework. The results show which classes of the framework are used most frequently in practice. This serves as input to Chapters 5 and 6.

This chapter was published in the Internation Journal of Human Computer Studies, and was co-authored by Guus Schreiber, Bob Wielinga and Marcel Worring (Hollink et al. 2004b).

2.1 Introduction

Recent advances in storage techniques have led to an increase in the amount of digital images all around the world. In addition, the growing accessibility of image collections has attracted more, and more diverse, user groups. These developments have heightened the need for effective image retrieval techniques.

Image retrieval techniques can be roughly divided into two areas: the traditional keyword-based approach and the relatively new area of content-based image retrieval. Although the latter in particular has seen much improvement in recent years (Smeulders et al. 2000), studies have shown that neither of these can answer the full range of user search questions. A disadvantage of the keyword-based approach is that the range of successful queries is limited to the interpretation of the indexer. Content-based retrieval systems, of which QBIC (Flickner et al. 1995) and Virage (Gupta and Jain 1997) are the first well known examples, are more flexible. However, they focus mainly on low-level features, while users search for high-level concepts (Eakins 2002). This is commonly referred to as the *semantic gap*.

The first step in resolving the mismatch between user questions and image retrieval techniques is to study the nature of user questions. Various authors have recognised the lack of knowledge about the way users search for images (Choi and Rasmussen 2002, Fidel 1997, among others). The aim of this chapter is to investigate user needs in image retrieval. We do this by asking two questions: (1) which categories of image descriptions exist? and (2) to what extent do people use each of these categories when formulating image queries? We present a framework for the classification of image descriptions in which we combine aspects from classification methods in

the literature. Categories in the framework can be used for both searching and indexing. A user specification is also added to the framework.

The resulting framework was used in an empirical study. Prior studies in this area (Armitage and Enser 1997, Heidorn 1999, Jörgensen 1998) have shown the importance of conceptual categories in image descriptions. However, more precise knowledge about the use of subcategories within the conceptual category is necessary to bridge the gap between user questions and image retrieval techniques. We used the categories of the classification framework to classify user image descriptions. Thirty participants were asked to read a text fragment and come up with a description of an image that illustrated the text. Then, they searched for a matching image using a web image searcher. The descriptions and queries were split into fragments and categorised in the framework. The results show how often each category is used in category search tasks.

2.2 Related work

To answer the question of what categories of image descriptions people use in the search process, we need a structure to categorise descriptions. Various classification schemes and methods can be found in the literature. A selection of these is discussed in the next subsections.

2.2.1 Theories

Erwin Panofsky (1962) developed a theory to structure content descriptions of images as early as 1962. He was an art historian who described three levels of meaning in Renaissance art: the 'pre-iconographical description', the 'iconographical analysis' and the 'iconological interpretation'. Sarah Shatford (1986) extended this model and showed its significance not only for renaissance paintings, but for all types of images. Based on Panofsky's three levels, Shatford categorised the subjects of pictures as 'Generic Of', 'Specific Of' and 'About'. At the GenericOf level, general objects and actions are described. Examples are woman, house or walking. The SpecificOf level describes individually named objects and events, such as the Eiffel tower or the fall of the Berlin Wall. The About level contains moods, emotions, abstractions and symbols, like happiness, justice or the iron curtain. Shatford also added four facets to each level: the 'who' facet, the 'what' facet, the 'where' facet and the 'when' facet. This resulted in a 3x4 matrix for the classification of the subjects of images. The so-called Panofsky/Shatford model has become a widely-used model for the classification of image descriptions and has been used by several researchers.

Jaimes and Chang (2000) classified image descriptions on the basis of the amount of knowledge required. They proposed a ten-level model for indexing based on both syntax and semantics. The higher the level, the more knowledge is needed to formulate a description. The first four levels are the so-called 'perceptual' levels. The first level is the type/technique level. It provides general visual information about the image. Examples of terms at this level are painting, drawing, photograph, black and white, colour and number of colours. The next three perceptual levels are based on the low-level features 'colour', 'texture' and 'shape'. A distinction was made between

(1) the characteristics of the image as a whole, (2) the characteristics of certain elements in the image and (3) the arrangement of the elements. The latter gives information about composition concepts, such as symmetry and viewing angle. No world knowledge is required to formulate perceptual descriptions.

The remaining six levels are 'conceptual' and can be seen as an extension of the Panofsky/Shatford model. The conceptual levels are divided into a generic level, a specific level and an abstract level, which directly corresponds to the division into GenericOf, SpecificOf and About. To each of these levels Jaimes and Chang added the distinction between descriptions about an 'object' in the image and descriptions about the 'scene' of the image as a whole. This makes six conceptual levels. General, specific or abstract world knowledge is required to formulate descriptions at the conceptual levels.

Eakins (1998) made a similar distinction, but focussed on *queries*, rather than on *indices* or the more general term *descriptions*. He identified three levels of image queries: queries based on primitive features, queries based on logical features and queries based on abstract features. Eakins based this arrangement on the distinction between primitive and logical features drawn by Gudivada and Raghavan (1995). Examples given by Eakins are "yellow and blue stars" (primitive), "a passenger train" (logical) and "pageantry" (abstract). Primitive queries correspond to the four lower levels in Jaimes and Chang (2000). The logical queries incorporate both general and specific descriptions in the model of Jaimes and Chang. Abstract queries are equal to the abstract descriptions as described by Jaimes and Chang.

2.2.2 Empirical studies

Although little empirical knowledge about classification of image descriptions is available, some experiments from various fields have added useful information to the theories above. Enser and McGregor (1992) analysed user requests submitted to the Hulton Deutch CD Collection. They represented the requests in terms of a 2 x 2 matrix of unique/non-unique, refined/unrefined queries. They found that 70 % of the requests was for a unique object. Later, Armitage and Enser (1997) used the Panofsky/Shatford model to find out which categories of image descriptions are used most by users of seven libraries. They found that the specific-who, generic-who and specific-where categories were used significantly more than average.

Jörgensen (1998) experimented with image descriptions from a cognitive psychological perspective. She analysed free text image descriptions and deduced 12 classes of image attributes that were used in the descriptions. She found that people describing images mainly used the 'object', 'people', 'colour' and 'story' classes. The latter contains descriptions of the event, the setting, activities and time aspects. The twelve classes of Jörgensen and the 2 x 2 matrix of Enser and McGregor were compared in the work of Chen (2001). Three reviewers classified image queries of 29 art history students using the two methods successively. Chen found that both were very well applicable and that the judgements made by the reviewers had a reasonable level of agreement: over 70 % of the classification judgements was agreed upon by at least two reviewers. The results

of the classification, however, differed from the original findings of Jörgensen and Enser and McGregor. The differences can be explained by different user characteristics, such as familiarity with indexing tools and different image domains.

Heidorn (1999) examined the mechanisms that people use in natural language to describe objects. He asked novices and experts to search for images of flowers by giving natural language descriptions. One of his results was that novices frequently used visual analogies, like "this looks like X". The results of these empirical studies all show the significance of conceptual descriptions in the search process.

2.2.3 Practical utilisation

Practical improvements in the field of indexing and searching have been accomplished through the introduction of the Dublin Core Metadata standard (Dublin Core 2006). This standard is used to add metadata to a wide variety of resources in a simple manner (Hillmann 2001). Similar to the Dublin Core categories, but focussing only on images, are the VRA Core Categories. The Visual Resources Association (VRA) formulated an "element set to create records to describe works of visual culture as well as the images that document them" (VRA 2002). The elements in this set describe various aspects of the *context* of images. Examples are author, date and title. The *content* of an image is described in a subject element. The VRA elements follow the 'dumb-down' principle (Hillmann 2001). In other words, the elements can always be reduced to Dublin Core elements. Some specific meaning may be lost but the value of the element is still generally correct. We come back to the VRA elements in Section 2.3.2.

2.3 A Framework for the Classification of Image Descriptions

2.3.1 Integration of methods into one framework

The classification methods discussed above all focus on different aspects of image descriptions. To answer the question "which categories of image descriptions exist?", it is necessary to take into account all these aspects. Therefore, we need to combine the components of different classification methods in one framework. A combined framework makes it possible to compare categories and to conclude on the importance of each category.

Some of the methods that are integrated in the framework are meant for indexing (Jaimes and Chang 2000, Shatford 1986), while others focus on searching (Armitage and Enser 1997, Eakins 2002). We combine the two by concentrating on image descriptions, which can be both search terms and indexing terms. This dual applicability is also found in the Dublin Core metadata standard and the VRA Core Categories. The work of Panofsky (1962) and Shatford (1986) is different from the other methods in that they structure *images* instead of *descriptions* of images. The framework that we propose categorises image descriptions.

Not only the *focus*, but also the *form* of the discussed classification methods varies. Jörgensen, for example, categorises descriptions into twelve topics, while Jaimes and Chang use ten succes-

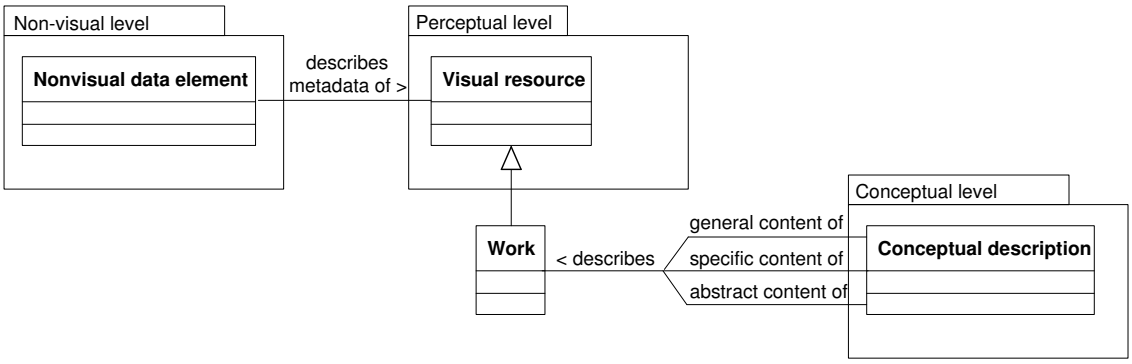

Figure 2.1 UML package diagram of an integrated framework for the classification of image descriptions.

sive levels. Variations in form make it difficult to see the differences and similarities between categories. To cope with this, we use the Unified Modeling Language (UML) to visualise the framework. UML (Booch et al. 1998) is a well-defined, standardised modeling language.

To realise the combination of different methods into one framework, we start from the similarities between the methods. Both Jaimes and Chang and Eakins have made the distinction between perceptual or low-level descriptions on the one hand and conceptual or logical descriptions on the other hand. Jaimes and Chang introduce an additional category: the non-visual information. This results in the three toplevels of the framework: the non-visual level, the perceptual level and the conceptual level (Figure 2.1). Each level consists of classes that represent categories of image descriptions. From here on we will refer to categories of descriptions as *classes*. A description of an image does not have to include all classes in the framework. Only the classes that represent the important features of the image are used.

The classes at the non-visual level are a subset of the VRA element set. They describe the context of an image, such as the ‘date’, ‘location’ and ‘creator’. The perceptual level contains the direct visual information about the image, like colours and shape. At the conceptual level the content of the image is described. The conceptual level is further divided into three sublevels: a general, specific and abstract sublevel. This distinction is made by several authors (Armitage and Enser 1997, Eakins 2002, Jaimes and Chang 2000, Panofsky 1962, Shatford 1986), although in different forms.

One of the biggest difficulties in image descriptions is the distinction between a *work* and the *physical instantiation* of a work. When describing a photographic image of a bronze sculpture, the describer can chose between two creators: the photographer and the sculptor. A digital image of a painting by van Gogh has two types of material: oil on canvas and pixels.

Shatford (1986) solves this problem by distinguishing between a work and a represented work. A represented work is the subject of a work. In the example above, the digital image would be the work and the Van Gogh painting would be the subject of this work. The painting itself also has a subject, e.g. sunflowers. The VRA takes a different approach. They provide a ‘record type’ element with two possible values: image or work. Work is the original work and image is the

work that represents the original. In the previous example, the digital image is the ‘image’ and the Van Gogh painting the ‘work’. Note that the use of the concept ‘work’ is the same as Shatford’s definition of a ‘represented work’.

The discussion between a work and a represented work is part of a bigger discussion about the occurrence of multiple copies of one original work. IFLA (1998), the International Federation of Library Associations and Institutions, distinguishes four entities:

Work An intellectual or artistic creation. The notion of a work is abstract.

Expression The specific intellectual or artistic form that a work takes each time it is realized.

Manifestation The physical embodiment of an expression, the materials.

Item A single exemplar of a manifestation. It is a concrete entity.

In our framework we make the distinction between a *work* and a *visual resource*. We define a work in accordance with the IFLA definition as a visual resource that is an intellectual or artistic creation. This also corresponds to the ‘work’ record type in the VRA element set and to the ‘represented work’ in Shatford (1986). In our framework `works` are a subset of the `visual resource` class. A visual resource is anything that is represented as or in an image. Visual resources can be works, analogue or digital representations of works, elements in images and items as defined by IFLA. The IFLA ‘expression’ and ‘manifestation’ classes are not incorporated in our framework. Since the form of the works in our domains is known to be ‘image’, an ‘expression’ class would be redundant. The ‘manifestation’ is partly covered in the `technique` class at the perceptual level.

The majority of the images in this study are digital representations of non-digital works. We expect users to search for both. When a user is searching for an image with a specific content, the search terms will refer to the original work. But when she is searching for an image to illustrate a website, characteristics of the digital representation, such as resolution and size, can be important.

Figure 2.1 shows that the non-visual and perceptual levels give information about visual resources. The resources are not necessarily works. The non-visual class `date` or the perceptual class `colour` are relevant for paintings as well as for photographs of those paintings. Conceptual information, however, is always about a work, since the subject of a represented work is always the same as the subject of the original work. We discuss the non-visual, perceptual and conceptual levels in detail in the following sections.

2.3.2 Non-visual level

At the non-visual level we are interested in descriptive information about the image. The information at this level is often called metadata. Hillman (2001) describes metadata as the “information that librarians have traditionally put into catalogs”. The information is about that carrier or medium of the image. This is in contrast to the perceptual and conceptual levels, where the information is about the content of the image.

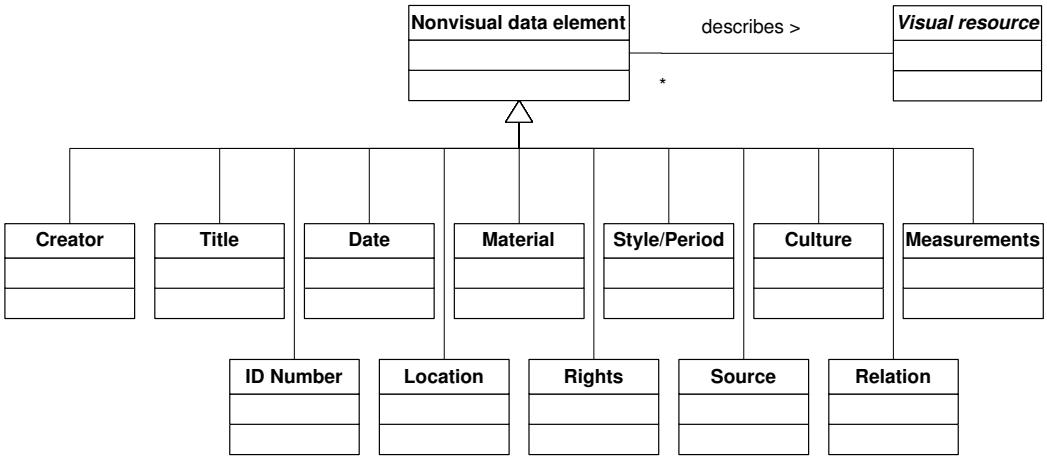

Figure 2.2 UML class diagram of the non-visual level.

We use two criteria to distinguish the non-visual level from the perceptual and conceptual levels. A description is non-visual if:

1. the information cannot directly be derived from the content of the visual resource.
2. the information is objective. It is not affected by interpretation.

Figure 2.2 shows the non-visual classes in a UML class diagram. The classes describe the context of a visual resource. `Date` contains dates associated with the image, such as `creation date` and `restoration date`. `Culture` specifies the culture or country from which the image originates. `Location` describes where the image is located, `rights` specifies who has the copyrights of the image and `source` contains the source of the information that is recorded about the image. `Relation` describes related works, such as other images of the same series. The remaining classes are self-explanatory.

The classes are a subset of VRA elements that meet the criteria for the non-visual level: they are objective and are not directly derived from the visual resource. VRA elements that are not included are Description, Subject, Type, Technique and Record Type. Description and Subject are captured in the perceptual and conceptual descriptions of the framework, Type and Technique are included as a class in the perceptual level and Record Type is covered by the distinction between a work and a visual resource.

2.3.3 Perceptual level

At this level we are interested in descriptions that are directly derived from the visual characteristics of the image. No knowledge of the world or the domain is required at this level. The first four levels of the model of Jaimes and Chang (2000) fall into this level, as well as the colour and visual element classes in Jörgensen (1998).

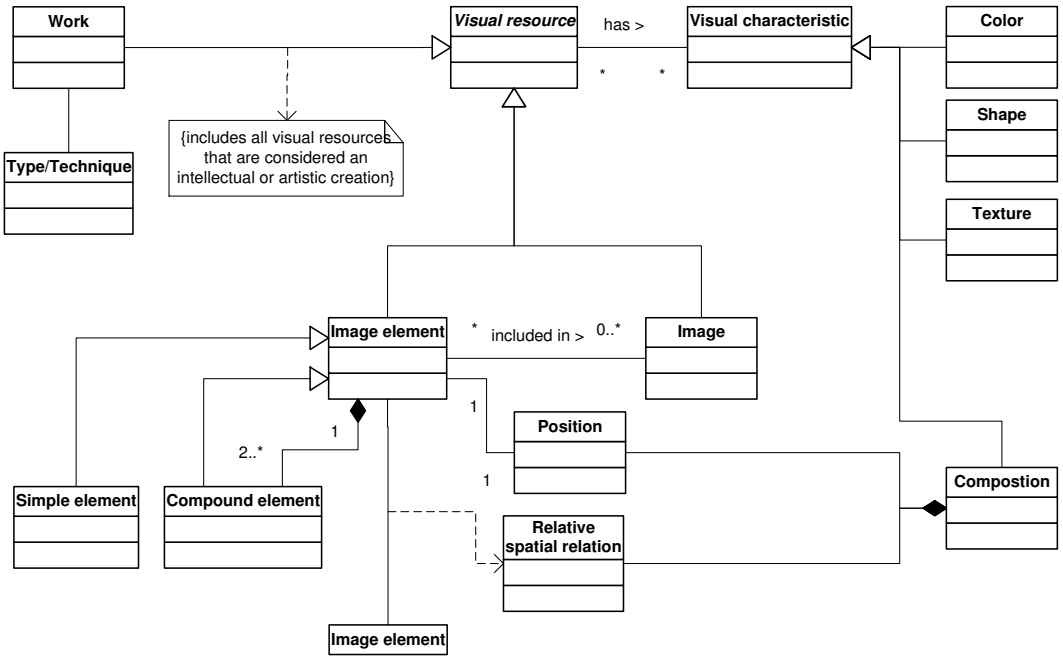

Figure 2.3 UML class diagram of the perceptual level.

The structure of the perceptual level is visualized in a UML class diagram (Figure 2.3) The `visual resource` class is in the centre of the diagram. A `visual resource` is either an `image` or an `image element`. In our case, the `image` class is defined as the collection of digital images we are working with. `Image elements` are parts of the digital image, or works that are represented by the digital image. Some `image elements` are `compound elements` and can be further divided into elements. The division into elements can continue several times, resulting in the 'Droste-effect', a Dutch phrase for infinite recursion in images. To illustrate the difference between an image and an image element, we may look at a user who is searching a database for a portrait by Rembrandt of his wife Saskia. The digital image in the database is the `image`. The original painting is an `image element` that is represented by the `image`. The painting is an `image element` that contains another element: Saskia.

Some visual resources are `works`, namely those who are considered an intellectual or artistic creation. In the example above, the painting by Rembrandt is a work, but the digital representation is not. Works can be described with a fourth characteristic: technique. The `technique` class is similar to the type/technique level in Jaimes and Chang (2000).

The `positions` of the elements in the image and the `relative spatial relations` between the elements together define the `composition`. `Composition` is one of the `visual characteristics` that a visual resource has. Other characteristics are `colour`, `shape` and `texture`. At this point, the relation between image descriptions and the low-level image characteristics they refer to, is extremely direct. Differences in opinion by users about the values of

the visual characteristics are negligible at this level. However, subjectivity cannot be completely avoided.

2.3.4 Conceptual level

The conceptual level gives information about the semantic content of the image. World knowledge is required for descriptions at this level. Experiments (Armitage and Enser 1997, Jörgensen 1998) show that people address this level frequently. As stated before, we divide the conceptual level into three sublevels, following the division made by both Shatford and Jaimes and Chang: a general, a specific and an abstract sublevel. For a complete description of an image, all three types of conceptual descriptions can be used at the same time.

General Concepts The general level is about generally known concepts. This sublevel requires only everyday knowledge of the world. An example of a description at this level is an ape eating a banana.

Specific Concepts This sublevel gives specific information about the content of the image. In contrast to the `general` sublevel, the objects and scenes are identified and named. Domain-specific knowledge is required at this sublevel. The ape in the example above can now be described as the old male gorilla Kumba, born in Cameroon and now living in Artis, a zoo in Amsterdam. The difference between general and specific concepts is not always clear. Armitage and Enser formulated this problem as follows (Armitage and Enser 1997): "an entity can always be interpreted into an hierarchy of related superconcepts and subconcepts; [...] it may not be obvious at what level one encounters the property of uniqueness." To differentiate between general and specific (or "unique") concepts, we use the *basic level categories* of Rosch. Basic level categories are the "level of abstraction at which [one] can obtain the most information with the least cognitive effort" (Rosch 1976). We classify descriptions that are more specific then the basic categories as specific and descriptions that are at the level of the basic categories or more general, as general.

Abstract Concepts At the abstract sublevel we add abstract meaning to the image. The knowledge used at this level is interpretative and subjective. To continue the above example, we could describe the content of the image as a species threatened with extinction.

The conceptual level is visualised in Figure 2.4. A `conceptual description` describes the content of a work. The description can be general, specific or abstract. This means that the content of one work can be described by three types of conceptual descriptions.

The `conceptual description` can be about a `scene` or about an `object` in the scene. Some objects are `compound objects` and can be further divided into parts, which are also objects. A `relation` can be defined between two objects. We follow Jaimes and Chang (2000) by making this distinction between scene and object descriptions.

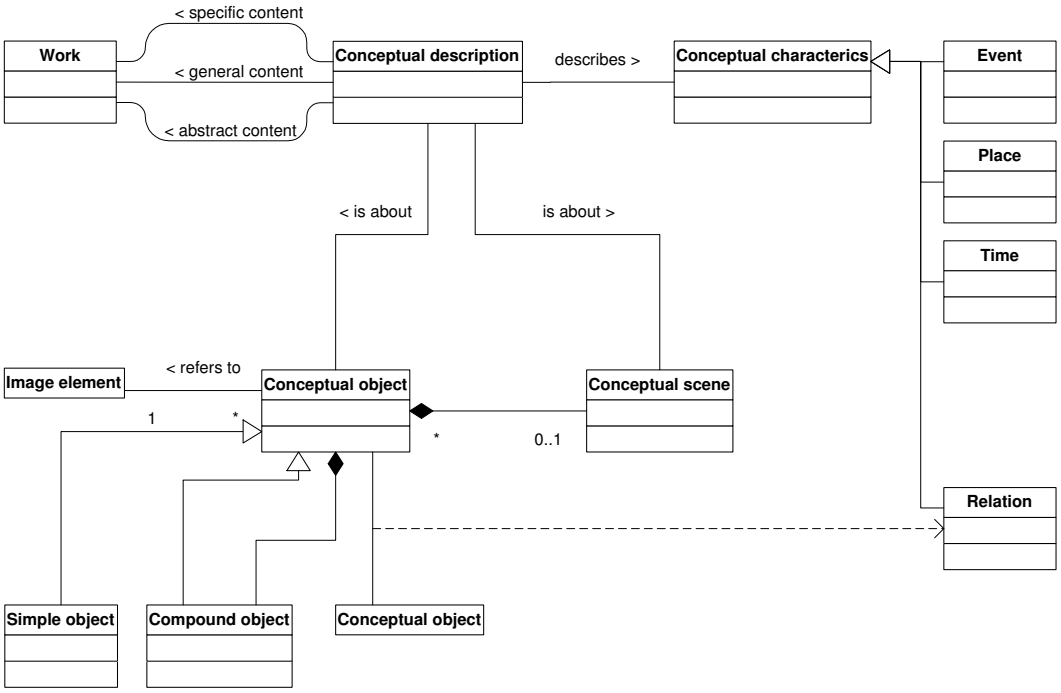

Figure 2.4 UML class diagram of the conceptual level.

A `conceptual description` consists of a set of conceptual characteristics: `event`, `place`, `time` and the `relation` between two objects. The `place`, `time` and `event` classes correspond to the 'where', 'when' and 'what' questions in Shatford (1986). The 'who' question is incorporated in the model as an instantiation of the conceptual `object` class.

The conceptual `object` class has a direct link with the perceptual level. A conceptual object refers to one or more perceptual `image elements`. In a description of a landscape, for example, a 'house' is a conceptual `object`. This object refers directly to a `visual element` in the image with the following `visual characteristics`: the `shape` is square, the `colour` is brown. In another image, the conceptual `object` 'house' refers to two `visual elements`: a red triangle on top of a brown square. 'On top of' is an example of a `relative spatial relationship` between two elements. The link between a conceptual `object` and a perceptual `element` is clearest at the general and specific sublevels. Abstract objects do not always have a perceptual counterpart.

2.3.5 Classification of users

The previous sections focussed on descriptions used in a search action. In this section the emphasis is on the searcher that formulates the descriptions. Characteristics of the searcher are important

for they can help to predict the classes of image descriptions that are used. We identify three user-related factors: (1) the image domain in which the user is searching, (2) the task the user is performing and (3) the expertise of the user. An overview of the user characteristics is shown in Figure 2.5.

Domain

The domain is the collection of images the user is searching in. Three characteristics of the domain are important for image descriptions: the breadth of the domain, the size of the vocabulary and the levels of expertise.

Breadth of the domain The breadth of the domain is the variability of the images within the domain. In narrow domains the variability is small and the techniques are similar for all images. In a broad domain the images vary in content and technique (Smeulders et al. 2000).

Size of the vocabulary The size of the vocabulary of a domain is the number of terms that are typically used in that domain. Also important is the ratio between domain-specific and general terms.

Levels of expertise Differences between levels of expertise within a domain vary across domains. Factors that define this characteristic of a domain are: differences in the amount of knowledge between experts and non-experts, the occurrence of intermediate levels, the effort needed to become an expert and the basic level of knowledge about the domain. In the domain of medicine, for example, there are large differences in the amount of knowledge between experts and non-experts. The effort needed to become an expert on a particular disease is large. Also, there are many intermediate levels of expertise. The group with the highest level of expertise consists of highly specialist doctors, followed by medical personnel that are not specialised in a particular disease. Then, there are patients suffering from the disease and finally laypersons who have never come across the disease. The basic level of knowledge is low: a large group of laypersons may never have heard of the disease. The football domain, on the other hand, has a broad user group with a reasonable amount of basic knowledge and a smaller group of experts. In this case, the difference between the levels of expertise is small.

Task

The task consists of three parts: the goal, the retrieval specification and the retrieval method.

Goal The goal is the reason behind the search, the activity that triggers the need to search for an image. Fidel (1997) organises searches along a spectrum starting with the ‘data pole’ and ending with the ‘objects pole’. At the data pole images are used as sources of information, while at the objects pole images are treated as objects. Images at the objects pole may be

used as decoration or to represent ideas. The illustration of a website is an example of a search at the objects pole. A student retrieving pictures of running horses to prove that all four legs of a galloping horse come off the ground simultaneously, is working at the data pole.

Retrieval specification The retrieval specification is the set of conditions the user expresses as input for the search. The conditions can refer to one or more aspects of the images. Smeulders et al. (2000) identify three types of searches: search by association, category search and target search. Users who search by association have no predefined idea of the content of the image and will not specify any conditions at the start of the search. In category search, the user has no specific image in mind but is able to specify requirements or conditions for the resulting image. The result will be the class of images that satisfy the conditions. This type of search is called search for an ill-defined target by Jaimes and Chang (2000). Target search is the type of search where a user aims at one specific image. The user will either use a copy of the image as input for the search, or will be able to express a highly precise set of conditions. This type is similar to the search for a defined target in Jaimes and Chang (2000).

Retrieval method The retrieval method refers to the tactics the searcher uses to express the retrieval specifications. Methods are linked to the three types of search as described by (2000). Browsing is a commonly used method for search by association. Textual queries such as free text or keywords with or without the use of operators like AND, OR and NOT and query by multiple example, are suitable for category search. Query-by-sketch is typically used in target search.

These descriptions of goal, specification and method are not complete. A more specific description is possible within a certain domain. Ornager (1995), for example, identified five types of query specifications in the domain of a newspaper image database: specific queries, general queries, queries in which the user tells a story and asks for multiple suggestions by the staff, queries in which the user asks the staff to suggest one most suitable image and fill-in-space queries where only the size and the broad category of the image are important. Such detailed typologies are only feasible and useful within relatively small domains. Since this classification of users is intended for the general domain of images, we define the categories of tasks broadly.

Expertise

The level of expertise of a user is defined by the amount of domain-specific knowledge she possesses. In a domain where several possible levels of expertise exist, a user has one of these levels. Expertise has implications for the retrieval method and specification. Vakkari (2000) analysed search tactics and search terms of students writing a research proposal. He found that as the participants' knowledge of the problem grew, the number of search terms increased. This increase was caused by an increase in related terms, narrower terms and synonyms. In his study, the

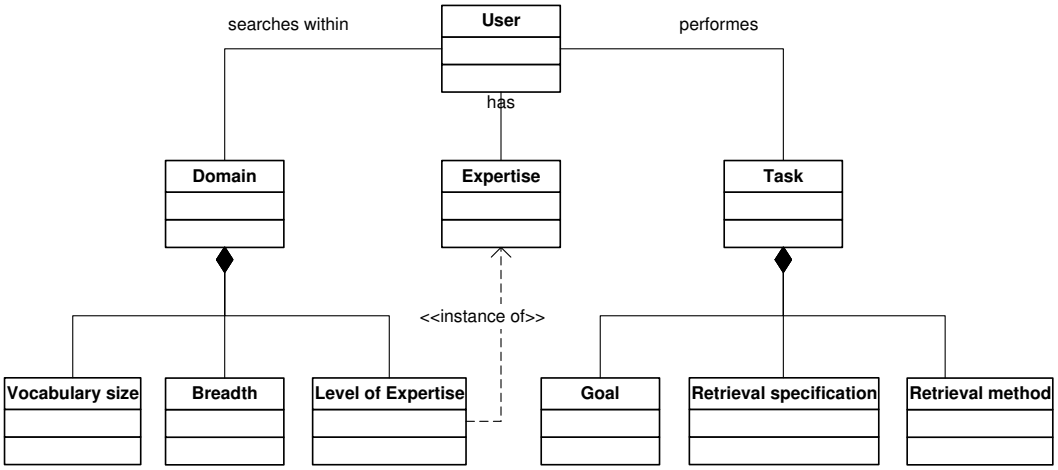

Figure 2.5 Classification of users.

use of broader terms decreased. Frost (1998) demonstrates that expert users prefer specific textual searches, while non-expert users, who have few domain-specific terms at their disposal, have more success with browsing and other visual search methods.

The study of user characteristics is a research topic in its own right. Studies like the ones by Vakkari and Frost show the relations between user factors. Also interesting is the influence of user characteristics on image descriptions. The impact of only one of the discussed factors, the domain, is addressed in the empirical study that is described in the next section.

2.4 User Study

2.4.1 Aim of the study

In Section 2.1 the following question was posed: “How much is each class of the framework used by people describing images?”. This question is relevant for the design of image retrieval applications. Information about the structure and components of user queries can help to improve query interfaces. In particular applications that use a structured vocabulary in which searchers or indexers describe images (e.g. Schreiber et al. 2001) can benefit from this information. Knowledge could be derived about what types of descriptions can be expected and what classes should be a part of the vocabulary.

In Section 2.2.1 we discussed the work of various authors who suggest that users prefer conceptual image descriptions to perceptual descriptions (Jaimes and Chang 2000, Jörgensen 1998). We are interested in the use of classes within these levels and in the use of classes in various situations. We investigated this in an experimental setting. Two different tasks and three different domains were included in the experiment. The tasks, a ‘describe’ and a ‘query’ task, are both

category search tasks. Category search has been described in Section 2.3.5 as the type of search in which "the user has no specific image in mind but is able to specify requirements or conditions for the resulting image. The result will be the class of images that satisfy the conditions." The three domains that are included in the study are represented by three texts that serve as context for the searches. The resulting image descriptions and queries were classified using the framework presented in the previous section. Specific questions that are addressed in the study are:

1. How much is each class used by participants performing a category search task?
2. What is the difference in the use of classes by participants between a 'describe task' and a 'query task'?
3. What is the difference in the use of classes by participants in different domains?

2.4.2 Methods

Thirty participants performed the overall task of illustrating a given text. This is a typical example of category search since several images can be suitable as an illustration. The participants were asked to read the text and form an image in their mind that could be an illustration of the text. Because the images are meant to adorn the texts, this is a task at Fidel's 'objects pole' (Fidel 1997). The overall illustration task was split into two subtasks. Firstly, they wrote down a free text description of the image. Secondly, they searched for a matching image using a keyword-based web image searcher[1]. They used a maximum of five queries, each containing one to a few words or phrases. The resulting free text descriptions and queries were classified in the framework, which led to an answer to the first question.

By comparing the descriptions to the queries we intend to find the difference between the semantically richer free text method and the limited but much used method of keyword-based search. Research has been done to investigate user image queries on the web (Goodrum and Spink 2001, Jansen et al. 2000), but it is still not clear how the terms used in web queries differ from the terms used in free text descriptions. Free text description is not common in current image retrieval systems because it requires the processing of natural language. But since it is a natural way for people to express their requirements, we chose this method to study user needs. From the comparison of the results we intend to answer the second question: "What is the difference in the use of classes by participants between a 'describe task' and a 'query task'?".

We repeated the experiment with three texts from different domains. The first text originated from a children's book, the second consisted of a few lines from a historical novel, and the third was a paragraph in a news item in a Dutch newspaper. The characteristics of these domains are not further specified in this chapter.

> **Text 1** "Not a day passed by without the squirrel taking a walk. He would drop himself from the beach tree on to the moss, or sometimes from the tip of a branch into the pond on the back of the dragonfly, which would take him in silence to the other side." (Translated from Dutch) [T. Tellegen. *Er ging geen dag voorbij*. Querido, Amsterdam, 1984]

[1] http://nl.altavista.com/image/

Text 2 "Evening after evening the pink and yellow in the air would melt together with the green of the fields, in houses the lights were turned on and eventually it grew silent everywhere, even if it was only for a moment, because then the birds started again as the first light broke the sky." (Translated from Dutch) [G. Mak. *Het ontsnapte land*. Atlas, Amsterdam, 2001]

Text 3 "Jorritsma, Kok and VVD state secretary Van Hoof of the Department of Defence spoke Saturday with the US ambassador Sobel. The Cabinet hopes to hear, this coming week, whether the Americans will give the Netherlands some more time. Kok said last Friday that the Netherlands will take a decision on the JSF-project after the elections, because the political power struggle in the country will be clear then." (Translated from Dutch) [Nederlands kabinet in de wacht gezet. *Volkskrant*, page 1, May 15, 2002]

To answer the third question, we compared the results of the three texts. Our aim was to get an idea of the influence of the domain on the use of search terms. The differences between the texts suggest possible differences in the results. The second text, which contains descriptions of colours and lights, may result in more perceptual descriptions than the other two texts. Text 3 contains specific names, which could lead to more specific descriptions than the other two texts. Ornager (1995) examined exactly the task for Text 3: the illustration of a newspaper item. Her findings that over half of the queries were specific, strengthen this idea.

Our approach differs from approaches in previous experiments (Heidorn 1999, Jörgensen 1998) in that the participants were not directly provided with an image. Instead, we provided the participants with text and asked them to imagine an illustration for the text. In this way an imaginary picture was used as the subject of description in the study. We chose this approach to reduce the bias due to visual cues in the image. When looking at an image, people tend to describe the aspects of the image that are clearly visible. These are not necessarily the same attributes that are important for the search task. A picture of the former Dutch prime minister Wim Kok wearing a bright red sweater will result in descriptions containing "red sweater", even though this is a trivial detail. By asking participants to describe an image that does not exist anywhere but in their minds, the descriptions will reflect those aspects of the content of the image that are considered significant by users. This is important for category search in particular, where the result of a search action is the set of images that display these significant aspects.

The following examples of descriptions and queries illustrate the use of various classes. Example 1 shows a description in which the technique class is used (e.g "drawn in detail"). The second and third examples contain specific terms (e.g "Chip and Dale", "Brussels"). The third example shows a typical utilisation of the abstract level in the phrase "narrow-minded dullness". All three examples include general level fragments (e.g. "squirrel", "trees", "men").

Example 1: free text description for Text 1 "A pencil drawn image or a still image from a cartoon. The image contains at least a squirrel and a tree with moss beneath it. The three and its branches are drawn in detail, a part of a pond is visible and a dragonfly."

Example 2: web query for Text 1 "Chip and Dale, trees, Walt Disney."

> **Example 3: free text description for Text 3** "A setting like we see in Brussels: Men walking in grey suits with hastily fastened ties. A narrow-minded dullness shows on the faces of the passers-by. Grey clouds lay over the city. A woman in a vivid red suit stands in the middle of the prevailing grey."

Some fragments in the descriptions were copied directly from the texts (e.g. "squirrel", "tree", "moss", "pond" and "dragonfly" in example 1, "trees" in example 2). To get an idea of how much the texts influenced the results, we scored the descriptions and found that 31 % was copied directly from the texts. This included plurals, singulars, diminutives and generalisations of words in the texts. The copied fragments were nevertheless included in the analysis for two reasons. Firstly, it is plausible that the participants picked these words from the text since they considered them to be relevant for their imagined image. Secondly, a certain influence of the experimental input on the results cannot be avoided. We measure this influence by comparing the use of classes in three different domains.

2.4.3 Subjects

The subjects were recruited from students of the University of Amsterdam and their family and friends. 15 Males and 15 females, aged between 15 and 56 (mean = 31) agreed to participate in the study. The majority was familiar with the internet: 96 % had been using the internet for more than a year and 85 % used it more than once a week. No evidence was found that gender, age, or use of the internet affected the use of classes of descriptions.

Of the 30 participants, three had read the book from which text one originated, one had read the novel from which text two came and five had read text three in the newspaper. None of them had a particular interest in or knowledge of one of the topics addressed in the texts, like a professional illustrator or political scientist would have had. Therefore, we consider none of the participants experts.

2.4.4 Analysis of the data

After collection of the descriptions, they were first split into fragments suitable for categorisation. This was done by parsing the sentences according to grammar. Fragments consisting of multiple visual cues about the content of the image were further split up into smaller fragments. Words that were not given in the context of a sentence were considered separate fragments. This process resulted in a set of 1151 fragments, each containing one or a few words. An example of a fragmented description of the second text is:

> "An image of | a village | at dawn, | the lights of most houses | are already on. | The vague contours of the houses | are still recognisable, | birds | fly | in the air. | The image | beams much warmth | because of the dark pink colour | of the air."

Subsequently, the fragments were assigned to a class in the framework. To preserve the meaning of the fragments, they were kept in the original context of the description. The data were then

normalised to compensate for differences in length of the descriptions between participants and between the two tasks: the number of occurrences of a class in a description was divided by the total number of fragments in that description. A fragment originating from a description containing ten fragments was counted as one tenth, while a fragment from a description split into five fragments was counted as one fifth. The total weighted number of occurrences of a class c in the study ($Count_c$) can be expressed as

$$Count_c = \sum_{i=1}^{180} \frac{n(d_i)}{N(d_i)}$$

where $n(d_i)$ is the number of occurrences of a class in a description i, $N(d_i)$ is the total number of fragments in a description i and 180 is the total number of descriptions as given by thirty participants performing two tasks on three domains.

Finally, the results were analysed by counting the number of fragments in each class. All analyses were performed on the weighted numbers.

The assignment of fragments to classes is a crucial part of the study. Although most of the assignments were straightforward, we cannot ignore the subjectivity of the assignment decisions. To estimate the proportion of the fragments that were sensitive to the personal interpretation of the assigner, two additional reviewers were asked to classify the data. To synchronise the classification decisions we used a set of guidelines for each part of the classification. The most important ones are the following. A complete copy of the guidelines can be found in Appendix A.

1. Consider a fragment to be specific only if it is more specific than Rosch's basic level categories (Rosch 1976).
2. Categorise all verbs as events, with the exception of forms of the verb 'to be'.
3. Consider a fragment to be abstract if the level of subjectivity is so high that differences in opinion about the interpretation are possible.

The classification decisions made by the two additional reviewers were compared to the classification used in this study. Cohen's kappa (κ)[2] was used as a measure of agreement between reviewers. We found a correspondence of 70 % (κ = 0.66) with the first additional reviewer and 75 % (κ = 0.70) with the second reviewer. This number is similar to levels of agreement between reviewers found by Chen (2001). To get more insight in the differences between reviewers, we examined the results further. The classification of a fragment consists of three parts: (1) the level of the fragment (non-visual, perceptual, general, specific or abstract), (2) the scope of the fragment, or whether it refers to the scene/image as a whole or to an object/element in the image and (3) the visual or conceptual characteristic (such as colour or shape, time or event) that can be assigned to the fragment. The first part had a correspondence rate of 83 % for the first reviewer (κ = 0.66) and 85 % for the second reviewer (κ = 0.67). The second part corresponded in 88 % (κ = 0.68) en

[2](κ) measures concordance between two raters using nominal data. κ varies between -1 and 1 (Brink, van den and Koele 2002). The degree of concordance is considered sufficient if κ is larger than 0.60

Table 2.1 Levels of abstraction in image descriptions in absolute numbers of occurrences and weighted percentages.

Level	Count	%
Non-visual	7	0.9 %
Perceptual	184	11.9 %
Conceptual	928	87.2 %
Total	1119	100.0 %

90 % (κ = 0.74) of the classifications. The third part reached a correspondence of 88 % (κ = 0.83) and 89 % (κ = 0.86).

These results show that in a relatively short time, a reasonable agreement between classifiers can be reached. Disagreements between classifiers were for the largest part contained in three classes. Firstly, the difference between the perceptual composition class and the conceptual relation class caused dissimilarity. Secondly, opinions about whether an object or scene could be specified as a place differed. Thirdly, differences between general and specific descriptions were not always clear. This is similar to conflicts that the reviewers in Chen (2001) had between the Unique and Non-unique categories.

2.5 Results

Analysis of the data resulted in a numerical overview of the levels and classes used by the participants. As expected, the conceptual levels were used most: 87 % of all elements were conceptual, 12 % were perceptual and only 0.9 % non-visual (Table 2.1). The near absence of non-visual expressions can be explained from the fact that we used imaginary images. The images described by the participants did not exist anywhere but in the participants' minds, so there was no author, date or rights. 32 Fragments consisted of personal remarks of subjects that did not concern the image directly, such as "ha ha ha", or "I see an image of ...". These were left out of the analysis, leaving 1119 valid fragments.

Table 2.1 shows absolute numbers of occurrences and weighted percentages of occurrences of the top-levels of description. Weighting was done to compensate for differences in length of the descriptions given by different subjects (see Section 2.4.4). The absolute numbers do therefore not always correspond to the percentages.

Because of the low number of non-visual descriptions, this level will not be discussed further. In Section 2.5.1 we report on the use of perceptual and conceptual elements by participants in this study. For this purpose, all descriptions are summed without differentiating between tasks or domains. In Section 2.5.2 we look at the differences in class occurrence between a 'describe' and a 'query task' and in Section 2.5.3 we discuss the differences between the three domains.

Table 2.2 Perceptual level: scope and characteristics of image descriptions in absolute numbers of occurrences and weighted percentages.

Scope	**Object**		**Scene**		**Total**	
Characteristic	Count	Weighted %	Count	Weighted %	Count	Weighted %
Colour	28	10.3 %	34	21.8 %	62	32.1%
Shape	2	1.2 %	2	1.3 %	4	2.5%
Composition	61	31.6 %	12	5.5 %	73	37.1%
Type/technique	5	2.5 %	40	25.9 %	45	28.4%
Subtotal	96	45.6 %	88	54.4 %	184	100.0 %

2.5.1 Use of classes of image descriptions in a category search task

Perceptual level

The perceptual level distinguishes between descriptions about the image as a whole and descriptions about elements in the image. In addition, five classes of descriptions are specified: colour, shape, texture, composition and technique. Within the perceptual level, the composition class was used most: 37 % (Table 2.2). These were mainly terms describing the relative spatial relationships between the elements. Examples are “an object (dragonfly) above an object (pool)” or “an object (squirrel) on another object (the back of a dragonfly)”. Colour accounts for 32 % of the perceptual elements and technique for 28 %. Shape was hardly used (3 %) and texture was not used at all. Descriptions of the image as a whole were used slightly more then descriptions of elements in the image, 54 % and 46 % respectively.

Conceptual levels

The general sublevel was the most frequently used level within the conceptual level (74 % of all conceptual descriptions). The remaining two sublevels, the specific sublevel and the abstract sublevel, were used less frequenlty, 16 % and 9 % respectively (Table 2.3). At the conceptual levels, objects were used more than twice as much as scenes (70 % and 30 %).

The descriptions were also categorised by the characteristics they describe. Not all fragments describe a characteristic; some mention an object or scene without specifying the event, place, time or relation. In fact, unspecified objects are the most frequently used class: 57 % of all conceptual fragments. Out of the four characteristics, event was used most (13 %), followed by place (12 %), time (6 %) and relation (2 %) (Table 2.3).

A log-linear analysis[3] was undertaken to test the existence of a relationship between the com-

[3]Log-linear analysis is a version of chi-square analysis that is used to analyse data containing more than two categorical variables. The purpose is to find out which variables are associated. Models of the data are created, ranging from a model of complete independence between all variables, through models that contain a subset of all possible relationships, to a model of complete dependence of all variables (the saturated model). The saturated model of a two-way table with variables A and B would be $\log \mu_{ij} = \lambda + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{AB(ij)}$, which represents the following

Table 2.3 Conceptual level: occurrences of conceptual classes in absolute numbers and weighted percentages.

Conceptual Sublevel		**General**		**Specific**		**Abstract**		Total	
Scope	**Characteristic**	Count	%	Count	%	Count	%	Count	%
Object	Event	98	7.4 %	2	0.1 %	8	0.7 %	108	8.3 %
	Place	22	1.9 %	9	0.7 %	0	0.0 %	31	2.6 %
	Time	3	0.2 %	0	0.0 %	2	0.2 %	5	0.4 %
	Relation	30	1.7 %	0	0.0 %	1	0.1 %	31	1.8 %
	Uncharacterised	381	40.1 %	114	14.2 %	24	2.8 %	519	57.1 %
Subtotal		534	51.5%	125	15.0%	35	3.8 %	694	70.3 %
Scene	Event	23	4.7 %	0	0.0 %	4	0.3 %	27	5.0 %
	Place	60	7.7 %	7	0.9 %	6	1.2 %	73	9.9 %
	Time	44	4.3 %	4	0.3 %	11	1.1 %	59	5.8 %
	Relation	3	0.3 %	0	0.0 %	0	0.0 %	3	0.3 %
	Uncharacterised	35	5.8 %	2	0.1 %	35	2.8 %	72	8.8 %
Subtotal		165	22.9 %	13	1.4 %	56	5.4 %	234	29.7 %
Total		699	74.4 %	138	16.4 %	91	9.2 %	928	100.0 %

ponents of a conceptual fragment. The three variables *sublevel* (general, specific or abstract), *scope* (object or scene) and *characteristic* (event, place, time, relation or unspecified) constitute a 3x2x5 contingency table (Table 2.3). The analysis fitted a model to the three variables that best explains the frequency data in the table:

$$[Sublevel\ Scope][Scope\ Characteristic]$$

This model shows dependencies between the level of abstraction and the scope of a description and between the scope and the characteristic of a description.The meaning of these findings is limited: the tests reveal nothing about the type and strength of the relations, nor about which categories within the related variables induce the relationship. However, based on this model and on the frequency of use of the classes in our experiment, we are able to propose the following hypotheses about co-occurrence of classes in image descriptions for category search tasks:

- We expect that abstract descriptions are more often about a scene than about an object in the image, while on average image descriptions are more often about an object than about a scene.
- We expect that time descriptions are always associated with a scene, while on average descriptions are more often about objects.

Further research is needed to test these hypotheses.

terms: expected value = constant + (row term) + (column term) + (association term). An alternative representation of this model is the shorter notation [AB]. Chi-square values are computed for all possible models to test which model fits the data best (Wickens 1989).

Table 2.4 Absolute numbers and weighted percentages of occurrences of (sub)levels of image descriptions in a 'describe' and a 'query task'.

Task	**'describe task'**		**'query task'**	
Level	Count	Weighted %	Count	Weighted %
Perceptual	147	17.3 %	37	7.0 %
General	450	63.4 %	249	67.5 %
Specific	57	8.2 %	81	20.4 %
Abstract	73	11.1 %	18	5.2 %
Total	727	100.0 %	385	100.0 %

2.5.2 Differences between a 'describe task' and a 'query task'

While we used all 180 image descriptions in the previous section, we will now look at the results for the two tasks separately. When comparing the 'describe task' and the 'query task', we see that the general level is the most frequently used level in both (Table 2.4). Differences can be seen at the remaining levels: in the 'query task' the descriptions contain more specific and less abstract and perceptual descriptions than in the describe task. A Chi-square goodness-of-fit test (Brink, van den and Koele 2002) showed that in our study there is indeed a significant difference in the distribution of descriptions over (sub)levels in the two tasks (χ^2 = 10.6, df = 3, $p < 0.05$). Thus, the data in our study suggest that people searching in a keyword-based search engine use more specific terms and less abstract and perceptual terms than people describing images in a more natural way. A possible explanation of this finding is that the various interpretations of abstract terms lead to low precision of the search results. Specific terms, on the other hand, lead to high precision of the results. People who have at least some experience in searching in keyword-based systems are aware of this effect and use it to enhance performance.

At the 0.05 significance level we did not find any evidence that the distributions of descriptions over characteristics and scope differ for the two tasks.

2.5.3 Differences between domains

A comparison of the results of the three texts shows the influence of the domain on the use of classes (Table 2.5). Descriptions of illustrations for the first text, a paragraph from a children's book, contain less abstract fragments than average (3 %). The third text, a newspaper item, contains more specific descriptions than average (36 %). The general level is the most used level for all texts, but descriptions for Text 3 hold less general descriptions than descriptions for Texts 1 and 2 (41 %). The percentage of approximately 12 % of perceptual fragments is similar for all texts. A chi-square goodness-of-fit test confirms a difference between the texts in the distribution of fragments over (sub)levels (χ^2 = 39.6, df = 6, p = 0.01).

In general, descriptions of objects or elements far outnumber descriptions of the scene or the image as a whole. Contrasting, descriptions of Text 2, a paragraph of a historical novel, consist

Table 2.5 (Sub)levels and scope of image descriptions in three domains in absolute numbers of occurrences and weighted percentages.

Domain		Text 1		Text 2		Text 3	
Level	**Scope**	Count	%	Count	%	Count	%
Element/	Perceptual	35	6.8 %	27	4.4 %	34	5.2%
Object	General	261	66.8 %	265	41.6 %	108	27.7 %
	Specific	12	4.5 %	1	0.1 %	112	34.7 %
	Abstract	9	1.6 %	2	0.8 %	24	7.6 %
Subtotal		317	79.7 %	195	46.9%	278	75.2 %
Image/	Perceptual	21	4.0 %	45	9.3 %	22	6.3 %
Scene	General	41	14.7 %	98	32.2 %	26	13.6 %
	Specific	0	0.0 %	5	1.9 %	8	1.7 %
	Abstract	7	1.6 %	36	9.6 %	13	3.1 %
Subtotal		69	20.3 %	184	53.0 %	69	24.7%
Total		386	100.0 %	379	100.0 %	347	100.0 %

of more scene descriptions than object descriptions (53 % and 47 % respectively). A chi-square test showed a significant difference at this point (χ^2 = 17.1, df = 2, $p < 0.01$). No indication was found that the distribution of descriptions over characteristics differs between the texts.

The differences seem intuitive. The first text is a simple story; mainly everyday knowledge is needed to understand it. People are not inclined to use abstract descriptions in this domain, which require more knowledge and interpretation. Text 3 describes a situation that has occurred in reality. This gives participants the possibility to use specific names and places in their descriptions. In Text 2, the high occurrence of scene descriptions could be caused partly by the high number of time specifications (13 % of all descriptions for Text 2). Time descriptions seem always to be associated with a scene.

The differences between the texts suggest a relationship between the domain and the classes of descriptions that participants use. The relationship, however, is not as straightforward as one might expect. In Section 2.4.2 we expressed the expectation that classes in the input texts would be reflected directly in the results; Text 2 contains perceptual classes which would result in perceptual descriptions, Text 3 contains specific classes which would thus result in specific descriptions. Yet the results show that the perceptual level does not differ significantly over the three texts. Text 3 did indeed receive more specific descriptions than Texts 1 and 2.

2.6 Discussion

The aim of our analysis has been to classify image descriptions and to study the use of each class in category search tasks. The outcome of an empirical study showed that the majority of the descriptions was conceptual (85 %). This is in line with findings of other researchers (Armitage and Enser 1997, Jörgensen 1998). Within the conceptual level, 74 % of the descriptions were

general, 16 % specific and 9 % abstract. Object descriptions were used twice as much as scene descriptions. Other frequently used classes were events and places at the conceptual levels and relative spatial relations at the perceptual level.

The experiment was designed to study category search. Other types of search, such as search by association or target search, may lead to other results. We expect that people performing a target search task will make more use of the non-visual level (which was not relevant in the present study). In addition, target search may lead to more `specific` descriptions, as the names of specific objects and scenes in the target image are known. Finally, we expect the perceptual level to be more important in target search, since the perceptual characteristics of the target image are known to the user.

Within our experiment the participants performed two category search tasks subsequently: a 'describe' and a 'query task'. For the design of search interfaces the results of the 'describe task' are more relevant. The free text descriptions in the 'describe task' were not biased by the limitations of existing search interfaces, while the queries in the 'query task' were.

We compared descriptions across three domains. Common findings in these domains were that (1) the `conceptual general` sublevel was the most frequently used level and (2) the perceptual level accounted for 11 to 14 %. The largest dissimilarity between the domains was found at the `conceptual specific` sublevel. This sublevel was significantly more important in the newspaper domain than in the other two domains. It seems that domains containing real-life images lead to frequent use of the `conceptual specific` sublevel.

Studies in other image domains may require specialisations of the framework. In the domain of art, for example, the style of a painting is an important image characteristic. As style is a subjective measure which needs interpretation and abstract world knowledge, it could be inserted in the conceptual level as an abstract characteristic which applies to works. Style can then be seen as the conceptual equivalent of the perceptual `type/technique` class.

In the process of using the framework, it appeared that guidelines are necessary about how to apply the framework. Some of the classes were ambiguous. The meaning of the classes `place`, `composition` and `relation` in particular was initially somewhat unclear to classifiers. A set of guidelines proved to be an effective way to come to a common understanding of the classes.

Comparing results from different experiments is inherently difficult. This is due to the different use of classes to categorise image descriptions. Still, we found some interesting similarities and contrasts between our results and results of experiments by Jörgensen (1998) and Armitage and Enser (1997).

Jörgensen used twelve classes to classify image descriptions. Two of these classes, the object and people classes, are the equivalent of the class of `object` descriptions in our framework. Here the two experiments show a similarity: Jörgensen's object and people classes are the largest of the 12 classes (Jörgensen 1998) and together account for 39 %, while none of the other classes exceeds 10 %; `object` in our framework is by far the most frequently used class and accounts for 50 % of all descriptions. We also found a discrepancy between Jörgensen's results and the present

study: in Jörgensen (1998) only 2 % of the descriptions are abstract, while the abstract level in the present study accounts for 8 %. A possible explanation is that the methods of collecting the image descriptions vary. In Jörgensen's experiment the users were presented with an image, while in our study they described an imaginary image. The latter may result in more abstract descriptions.

A comparison of our results with the work of Armitage and Enser (1997) also showed some similarities and differences. They used the Panofsky/Shatford model (Shatford 1986). The `object` class in our framework is the equivalent of the 'who' question in the Panofsky/Shatford model. Both studies show that this class of descriptions is used most frequently. The main difference between the results of Armitage and Enser and the present study can be found in the use of the specific level. In the study of Armitage and Enser the specific level is the most frequently used level, while in the present study the general level is by far the most frequently used level. The difference could in part be due to different domains that are used in the studies. We saw that the number of specific descriptions was higher for the newspaper text than for the novel and the children's book. Enser studied domains that are similar to the newspaper domain: seven libraries with collections containing photos and films about geography, film and television and (local) history. The images depict real scenes and objects, in contrast to children's books and novels, which contain fictive images. Another possible explanation for the difference are the varying methods. Armitage and Enser used queries that were put by users of the libraries. It is possible that the library users used specific terms because they knew from experience that specific queries are effective for retrieval. Note that the results of the 'query task' in our study also resulted in more specific descriptions than the 'describe task'.

Insight into the use of classes of image descriptions is useful in the design of image retrieval systems. In spite of improvements in the field, the discrepancy between the concepts of users and the possibilities of retrieval systems still exists. This discrepancy is referred to as the semantic gap (Smeulders et al. 2000). The aim of this chapter was to contribute to the knowledge about one side of the gap, namely the concepts of users. The next step will be to introduce a link between these concepts and technical possibilities. Such a link exists between the conceptual level and the perceptual level. A `conceptual object` refers to one or more `perceptual elements`. The results of the present experiment show that the conceptual object class is typically used in image descriptions. A perceptual element is described by colour, shape and texture, which are characteristics that can be used in automatic retrieval of images. This seems a logical point to look for a link between user concepts and retrieval techniques.

Acknowledgements

We would like to take this opportunity to thank Suzanne Kabel and Vera Hollink for additional classification of the image descriptions, Esther Bisschop, Janneke Habets, Sharon Klinkenberg, Bas van Nispen, Menno Scheepens and Marjolein van Vossen for collection of the data and Giang P. Nguyen for comments.

3

Evaluating the Application of Semantic Inferencing Rules to Image Annotation

In this chapter we investigate the possibilities of a direct link between the low-level visual features of an image and the high-level domain concepts that humans use. We describe an approach in which domain experts formulate rules that link visual features to domain concepts. By comparing the performance of this approach across types of image content and across domains, we determine characteristics of a domain which indicate whether a direct link is feasible. The results give insight in when the semantic gap limits retrieval, and help interpret the results of Chapters 4, 7 and 8.

This chapter was co-authored by Suzanne Little and Jane Hunter and published in the proceedings of the International Conference on Knowledge Capture (Hollink et al. 2005c).

3.1 Introduction

Semantic annotation of visual resources is essential to ease the discovery of the rapidly increasing quantity of digital visual content. Such descriptions enable sophisticated semantic querying of media in terms familiar to the user's domain whilst also ensuring that the information and knowledge have a much greater chance of being discovered and exploited by services, agents and applications on the web. Because of the quantity and complexity of visual data, manual annotation is slow, expensive and highly subjective. Despite advancements in the field of image analysis, the automatic generation of high-level semantic annotations of images remains a significant challenge.

Earlier research (Hunter et al. 2004) developed a semi-automatic, user-assisted approach to generating ontology-based annotations of image regions from low-level, automatically extracted features. This prototype enables experts to define rules specific to their domain, which map particular combinations of low-level visual features (colour, texture, shape, size, etc.) to high-level semantic terms defined in their domain ontology. These semantic inferencing rules capture a domain expert's understanding of how low-level features are related to ontology terms. The rules are recorded in an XML-based format and can be shared, collaboratively modified and annotated as the domain understanding shifts and changes.

To overcome the difficulty that domain experts face when developing complex rules in XML format using unfamiliar terminology, a visual interface called Rules-By-Example (RBE) was de-

veloped. To give an example, an biologist labeling cellular images to enable the search and retrieval of particular types of cellular components, may define the following rule:

example rule

If ((the component is dense) and (the dominant colour of the component is 100))
then (the component is a mature granule).

The system assists users to construct rules with palettes of example colours, shapes defined using drawing tools and example regions within the media collection.

The linking of low-level image data to high-level domain concepts is challenging due to what Smeulders et al. (2000) call the semantic gap. The question arises as to what are the conditions under which our approach can successfully establish such links and apply them to image annotation. The approach of RBE and semantic inferencing rules was previously evaluated in the domain of fuel cell microscopy (Little and Hunter 2004) where a small study demonstrated promising results. We extended the system to adapt it to the more complex domain of pancreatic cells. By comparing the characteristics of the two domains and examining the results and types of rules produced, we aim to determine the characteristics of domains which may benefit from semantic inferencing rules and the Rules-By-Example system.

In this chapter we describe the system architecture and the changes made to the RBE and semantic inferencing system in order to support this new complex domain of pancreatic cell analysis. The cell domain is presented together with the vocabularies that were used. We describe the rules that were defined, determine the accuracy of these rules and discuss the characteristics of domains in which this approach is likely to succeed.

3.2 Related Work

A number of research efforts have investigated the use of automatic recognition techniques to extract low-level visual and audio features which together can be used to generate semantic descriptions of multimedia content. These include statistically-based machine-learning methods (e.g. Adams 2003, Chang et al. 1998, Naphade and Huang 2001, Zhao and Grosky 2002) which first manually annotate sample sets and from this generate factor graphs, statistical models or other indexing techniques for the larger collection. Marques and Barman (2003) integrated ontologies into a machine-learning based approach to semantically annotate images.

Overall, the use of machine learning techniques to bridge the semantic gap provides a relatively powerful method for discovering complex and hidden relationships or mappings. However, the ‘black box’ method often employed can be difficult to develop and maintain because its effectiveness depends on the design and configuration of multiple variables and options. The relationships discovered between low-level media features and semantic descriptions remain hidden and cannot be examined or manipulated by the domain expert. In addition, extensive, detailed and specific training corpora are required to ensure optimal performance. These can neither easily be adapted to new domains nor incorporate new content or knowledge.

Methods for linking visual thesauri or ontologies to multimedia have been developed by Hoogs et al. (2003) and Tansley (2000). These methods utilise the relationships described by the ontology or thesaurus to enable more complex semantic queries across collections of annotated media and to infer new information. However, the difficulty of forming the relationships between media terms and domain terms still remains.

Rules have long been used as a means of capturing expert knowledge (Buchanan and Shortliffe 1984, Sowa 2000). If recorded in an open and transparent fashion, they are able to clarify the understanding of a domain's paradigm and act as a catalyst for discussion and exchange. The semantic web initiative includes a layer for logic, reasoning and rule processing. The standards (XML, RDF, OWL, RuleML, SWRL) and processing tools (CWM, Mandarax, JESS) that are emerging to supply this layer are intended to provide open, interoperable formats for the exchange, discussion and application of data, ontologies and rules. For example, Hatala and Richards (2003) have used ontologies in combination with rules to improve metadata for learning objects by suggesting relevant values.

We believe that our approach overcomes some of the limitations in existing image annotation approaches, such as the difficulty of determining the distinguishing features and adapting to different domains. The resulting semantic inferencing rules are a form of knowledge in themselves and can be discussed, annotated, shared and applied as the user directs. We accomplish this through the complementary use of semantic web technologies and an interface which allows domain experts to intuitively and interactively develop and define semantic inferencing rules in an interoperable, machine-understandable and shareable format.

3.3 The Domain of Pancreas Cells

At the Institute for Molecular Bioscience of the University of Queensland, the Visible Cell project aims to heighten the understanding of processes in mammalian cells. One of the goals of this project is to create a three-dimensional image of a cell. To this end, pancreatic cells were cut into 400 nanometre thick slices and each slice was studied by electron tomography (Marsh et al. 2001). Molecular biologists segmented and annotated the images of these slices by drawing lines around each cellular component (Figure 3.1).

480 Of these images were combined into a single, high-resolution three-dimensional reconstruction of a 3.1 x 3.2 x 1.2 μm^3 volume in a pancreatic cell. Figure 3.2 illustrates the spatial layout of a number of important cellular components including the Golgi apparatus, endoplasmic reticulum, mitochondria, ribosomes and different types of vesicles. A new high-throughput microscope that is capable of producing even larger numbers of images will soon be employed, heightening the need for automatic segmentation and annotation. The present chapter aims to contribute to the (partial) automation of the annotation process. Automatic segmentation is outside the scope of this chapter.

Segmented two-dimensional images were used as input to our RBE system. Each segmented region depicts a cellular component. The components that are visible in a cell all have distin-

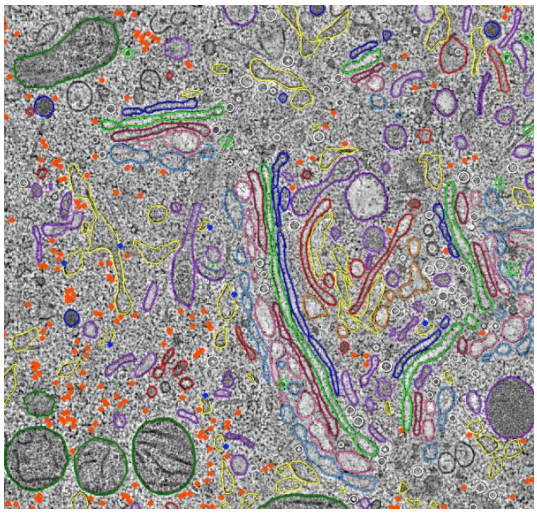

Figure 3.1 Segmented image of a pancreatic cell.

Figure 3.2 Three-dimensional image of a pancreatic cell.

guishing combinations of textures, colours, sizes and other features. We used the Matlab Image Processing Toolbox to extract these features. The toolbox provides a number of built-in functions for feature extraction. Custom routines for extracting additional features can be constructed relatively easily.

The pancreatic images differ in a number of ways from the images in the previous domain of fuel cell microscopy. In the pancreas domain a larger number of classes has to be identified, there is less visual distinctiveness between objects of different classes and there is less visual uniformity of objects of the same class. Because of this, the rules in the pancreas domain are more complex than in the fuel cell domain.

3.4 Vocabularies

In order to produce semantic annotations, we used existing vocabularies to represent the domain of pancreatic cells, the visual image features and the semantic inferencing rules. In this section we will describe the vocabularies that we used together with our experiences working with the molecular biologists to define, create and extend them.

3.4.1 A vocabulary for the pancreas domain

Our domain involves semantic descriptions of cellular components. The Medical Subject Headings thesaurus[1] (MeSH) is used for indexing and searching biomedical and health related documents. Since MeSH contains a hierarchy of cellular components we decided to reuse this large existing thesaurus, rather than build our own. We used the version of Van Assem et al. (2004), who translated MeSH from the native format to RDF. In meetings with molecular biology experts, we established the terms that they use to describe the cellular components in this collection of pancreatic cell images and the mappings between these terms and MeSH concepts. Direct mappings were not always possible due to the fact that the experts use functional descriptions such as “vesicle carrying cargo”, or visual descriptions such as “small tubular vesicles”. In these cases we extended the MeSH thesaurus with subclasses of the class vesicle.

3.4.2 Multimedia ontology

The Mpeg-7 ontology is used as a vocabulary for the visual image features. Mpeg-7 is a standard for describing multimedia content published by the Moving Picture Experts Group. It provides Multimedia Description Schemes, a Description Definition Language and tools that support generation of Mpeg-7 descriptions (Martínez 2001). The Mpeg-7 OWL ontology as published by Hunter (2001) includes low-level visual properties such as colour, shape and motion.

In various meetings with the domain experts, we asked them to identify those visual characteristics that distinguish different cellular components. We used the Matlab Image Processing

[1] http://www.nlm.nih.gov/mesh/meshhome.html

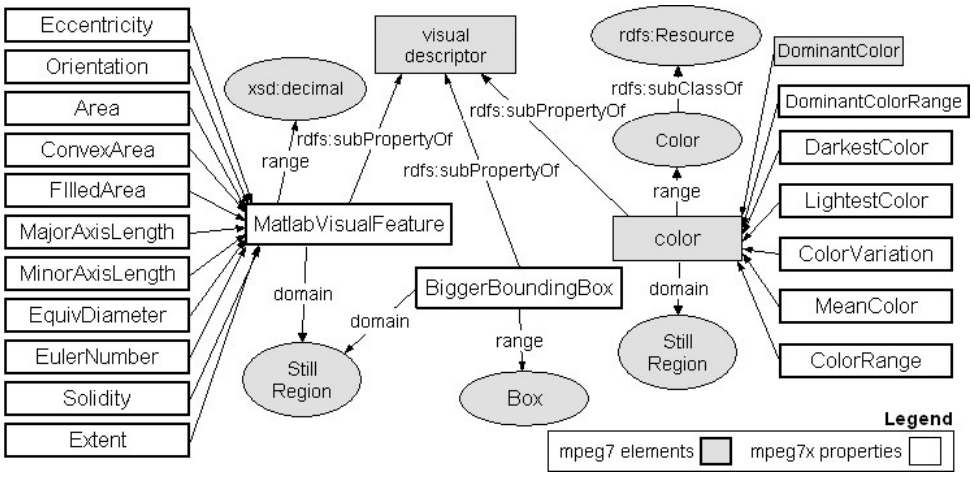

Figure 3.3 Mpeg-7 Visual Descriptor Colour.

Toolbox to extract these visual characteristics for each region. However, Mpeg-7 does not contain classes for all these detailed visual features. Therefore, we extended Mpeg-7 with 14 Matlab built-in image features (e.g., area, eccentricity) and 10 custom image features that we constructed with the Matlab Image Processing Toolbox (e.g., colour range, mean colour) (see Figure 3.3). All Matlab image features were represented as subproperties of the existing Mpeg-7 property visualDescriptor. The custom image features all concerned colour characteristics and were added as subproperties of mpeg7:colour.

Many of the Matlab visual concepts are too specific to be comprehensible to anyone but image analysis specialists. We cannot expect biologists to understand what 'eccentricity' is or what values the property 'density' may take. Instead of these specialist terms, the molecular biologists use more commonly known terms like 'long', 'round' and 'close' to describe visual characteristics. We defined rules to translate low-level terms such as 'eccentricity' and 'density' to more familiar commonly-used intermediate-level terms like 'long' and 'close'. Abella and Kender (1999) researched such links between low-level features and commonsense terms. They found, for example, that humans consider two objects to be close to each other if the bounding boxes of two objects multiplied by 1.6 overlap. We used this to construct the following rule, in which the 'bigger bounding box' is defined as 1.6 times the original bounding box:

bigger bounding box

If (the bigger bounding box of region1 overlaps with
the bigger bounding box of region2)
then (region1 is close to region2).

Additional intermediate-level terms that we defined in this way include: `long`, `solid`, `irregular`, `round`, `dense` and `touching`. These were added to the Mpeg-7 ontology in the same way as the low-level Matlab image features, as subproperties of existing Mpeg-7 properties.

3.4.3 Rule language

We used RuleML to represent both the rules that relate low-level visual features to intermediate-level terms and the semantic inferencing rules that experts defined to annotate cellular components. RuleML aims to provide a shareable, XML-based rule markup language for rule storage, interchange, retrieval and firing/application (Boley et al. 2001). Using this format ensures that our rules are machine readable and interoperable with existing tools and standards. A proposal for a new rule language that combines RuleML and OWL has recently been submitted to the W3C as the Semantic Web Rule Language (SWRL) (Horrocks et al. 2004). We plan to upgrade to SWRL once it stabilises and tool support increases, since it is more expressive than RuleML. The biggest improvement from our point of view is the inclusion of numerous built-in relations. For readability, we use the SWRL informal syntax rather then the lengthy XML syntax of RuleML in the rule examples throughout this document. A typical example of a rule in our domain is the rule below which is applied to recognise mature granules in an image.

Mature granule

mpeg7:StillRegion(region) $\wedge$ mpeg7x:Dense(region) $\wedge$
mpeg7:DominantColor(region, col) $\wedge$ swrlb:lessThan(col, 100)
$\rightarrow$ mpeg7:Depicts(region, mesh:MatureGranule)

This rule states that a region depicts a mature granule if it is dense and its DominantColor value is less than 100. The prefixes `mpeg7`, `mpeg7x` and `swrlb` indicate terms from the Mpeg-7 ontology, our extensions to the Mpeg-7 ontology and SWRL built-ins, respectively.

3.5 Semantic Inferencing and the Rules-By-Example Interface

As the previous section demonstrates, semantic inferencing rules to relate low-level image features to high-level semantic annotations can be complex, require understanding of specific media feature definitions and are formatted in XML. In contrast, domain experts generally have only a basic understanding of rule structure and its application, often have limited knowledge of media terminology and rarely enjoy constructing XML syntax. To overcome this, the Rules-By-Example system exploits domain ontologies and multimedia ontologies. It facilitates the construction of intermediate rules to describe media features in terms more commonly used by domain experts (long, dark, etc.) and contains example-based definitions of media features.

3.6 Previous Work

Figure 3.4 illustrates the components of the Rules-By-Example (RBE) system. The application utilises the Mpeg-7 descriptions and ontology and the domain ontology to build an interface which incorporates familiar, semantic terms. Mpeg-7 based descriptions of the regions within the image are loaded into the RBE application and the user is able to specify values for visual features

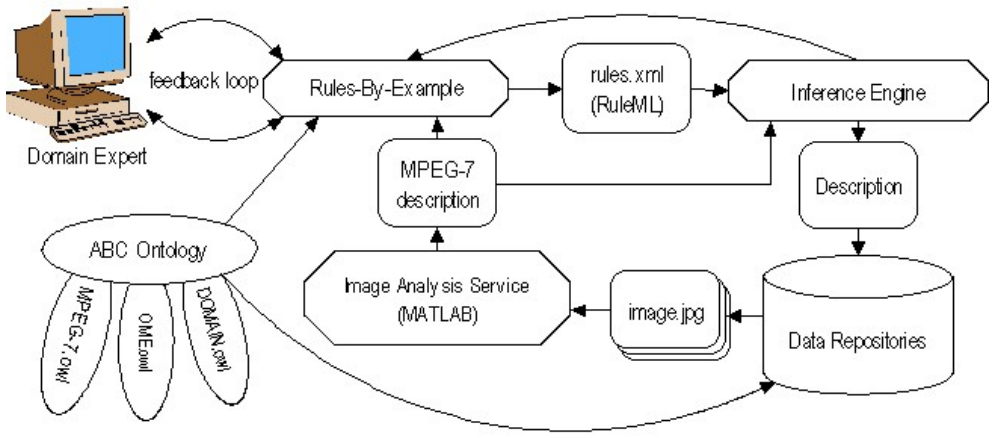

Figure 3.4 Architecture for the RBE System.

from palettes of sample colours and textures, drawing tools or by specifying regions from sample images selected from the data repository. This method of defining the rules, 'by example', is more intuitive and reduces the prior knowledge and understanding required to build semantic inferencing rules. For example, the user is able to drag-and-drop colour selections as opposed to entering RGB definitions. More information about its implementation can be found in Little and Hunter (2004).

As the user is constructing a rule, the system is able to evaluate it against a sample set of manually annotated images and present the number of currently matching regions. This simple feedback enables the user to determine when a rule may be most accurate or when a rule has become too specific. Once the user is satisfied with a rule, it is saved in RuleML format, augmented with MathML where necessary to describe mathematical relationships, and saved to an XML database. This can then be made available to collaborators over the web for discussion, re-use and refinement as a result of application to their own collections of images. A complete set of metadata describing the rule, including evaluation results and the data set used in development, is also recorded to ensure that the provenance of the process is well documented.

Overall, the Rules-By-Example interface allows the user to quickly develop, apply and refine highly complex rules while reducing the need to understand low-level Mpeg-7 terms or values. In addition, it enables users to direct the system in such as way that it focuses on the objects, regions or distinguishing features of highest priority – in contrast to traditional approaches which require pre-selection of important features. The rules themselves are recorded in a flexible, interoperable format and together with their provenance metadata can be easily distributed, discussed and modified.

3.6.1 Application to pancreatic cell images

Since the prototype RBE interface was developed for the fuel cell domain, changes were required to apply it to the new pancreatic cell domain. Firstly, the fuel cell ontology was replaced by

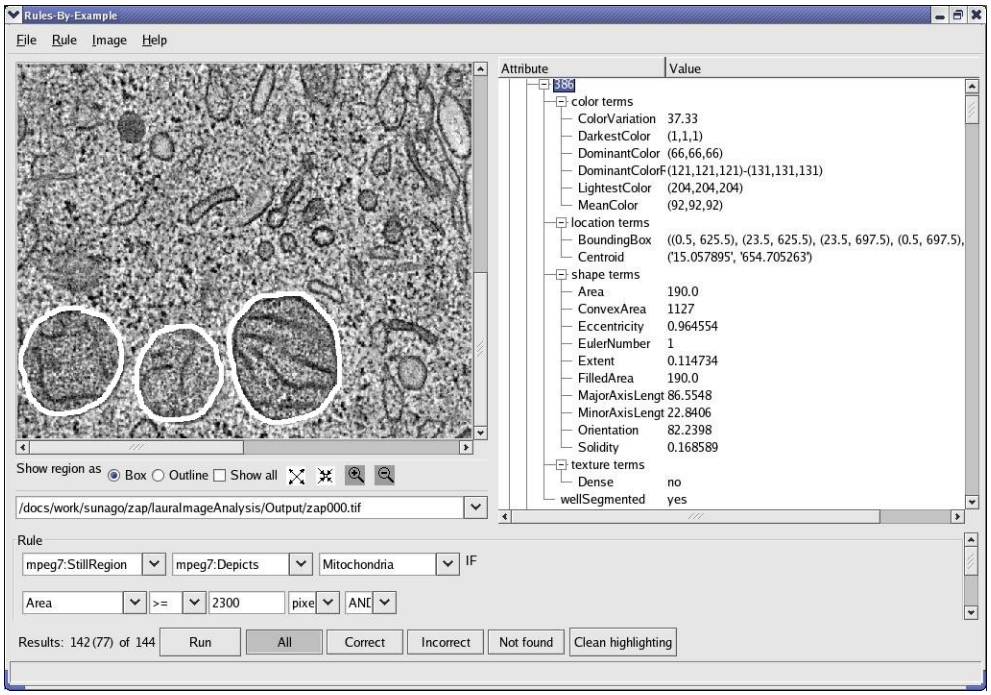

Figure 3.5 Screenshot of the RBE System, showing three mitochondria that are manually segmented by a domain expert, and the first part of a rule that identifies mitochondria based on the visual property `area`.

the MeSH ontology. Secondly, changes were made to the data access module to connect to the new database and schema used for the pancreatic cell images. The most significant change was required to accommodate the greater complexity of the pancreatic cell images. This required the definition of more intermediate rules than were required by the simpler fuel cell images. Changes were made to the way the available media properties, including the extensions we made to Mpeg-7, were loaded into the interface and displayed with the image's metadata. Figure 3.5 shows a screenshot of the RBE interface, highlighting regions in the pancreas image which match the current rule.

The application and processing of the rules remains a bottleneck for the system. Due to the limitations of RuleML in defining mathematical relationships and the current limitations of the inferencing engine used in this prototype (Mandarax), pre-processing of data by Python scripts is required to apply relationships such as less than, equal to and greater than. Processing of the rules involved a combination of Python scripts, complex MySQL queries, pre-application of intermediate rules and processing through the Mandarax inferencing engine. This series of steps severely limited the accuracy and responsiveness of the dynamic feedback component of the RBE interface and may have had a negative impact on rules developed by the domain expert.

A primary goal of the RBE system was to maintain as much domain independence as possible by enabling different domain ontologies to be easily plugged in. It was a simple process to replace

the existing fuel cell ontology with the MeSH ontology to provide the semantic terms for the rules. The increased complexity of the cell images required extensions to provide greater support for intermediate rules. The issue of processing speed that limits the dynamic response is a critical one. Currently, a number of alternative tools and standards, such as SWRL translation through XSLT to CLIPS (JESS) or Prolog are being investigated to overcome this.

3.7 Evaluation

3.7.1 Setup

Thirty pancreatic cell images, all slices from one cell, were used to evaluate the system. The regions had previously been segmented manually by domain experts into a total of 7580 regions. A domain expert (molecular biologist) used the Rules-By-Example interface to create rules for the annotation of cellular components occurring within the images. Application of the rules generated annotations for each segmented region in the images. The correctness of the annotations was determined by comparing them with manual annotations that had been specified at the time of segmentation.

The biologists' feedback on the RBE interface was, on the whole, positive. They noted, however, that the system would only be helpful if combined with accurate automatic segmentation. A project on automatic segmentation of pancreatic images is currently underway.

3.7.2 Rules formulated by a domain expert

Five rules were formulated by the domain expert to identify five classes of cellular components: Golgi stack; endoplasmic reticulum; mitochondrion; ribosome; and mature granule. The Golgi apparatus consists of a series (a 'stack') of long thin components alongside each other. Looking at the components separately would not distinguish them from components of other types, but when two or more of them occur close to each other, it is a very strong indication that they are part of a *Golgi stack*. Hence, the rule to recognise a Golgi stack is: a region depicts a Golgi apparatus if its eccentricity is greater than 0.98 and it is close to a region with an eccentricity greater than 0.98.

Golgi apparatus

mpeg7:StillRegion(region) $\wedge$ mpeg7x:eccentricity(region, ecc) $\wedge$
swrlb:greaterThan(ecc, 0.98 $\wedge$ mpeg7x:close(region, region_y) $\wedge$
mpeg7x:eccentricity(region_y, ecc_y) $\wedge$ swrlb:greaterThan(ecc_y, 0.98)
$\rightarrow$ mpeg7:Depicts(region, mesh:GolgiApparatus)

The endoplasmic reticulum (ER) is an irregularly shaped component that can be distinguished by the fact that ribosomes are attached to it. However, the images in our collection are slices of a three-dimensional cell and the connection with a ribosome is not necessarily visible in every slice. Therefore, the rule for ER is: a region depicts an endoplasmic reticulum if the region is touching a

ribosome or the region is adjacent to a region in a consecutive slice, which is touching a ribosome. We used Python scripts to calculate if two regions were ‘adjacent in consecutive slices’.

Endoplasmic reticulum

mpeg7:StillRegion(region) ∧
(touching(region, region_a) ∧ mpeg7:Depicts(region_a, mesh:Ribosome)) ∨
(z_adjacent(region, region_b) ∧
touching(region_b, region_c) ∧ mpeg7:Depicts(region_c, mesh:Ribosome))
→ mpeg7:Depicts(region, mesh:ER)

The rules for mitochondria and ribosomes are shown below. Section 3.4.3 describes the rule for mature granules.

Mitochondrion

mpeg7:StillRegion(region) ∧ mpeg7x:shape(region, mpeg7x:solid) ∧
mpeg7x:eccentricity(region, eccentricity) ∧
swrlb:greaterThan(eccentricity, 0.9) ∧ mpeg7x:area(region, area) ∧
swrlb:greaterThan(area, 2300) ∧
mpeg7:DominantColor(region, dominantColor) ∧
swrlb:greaterThanOrEqual(dominantColor, 105)
→ mpeg7:Depicts(region, mesh:Mitochondrion)

Ribosome

mpeg7:StillRegion(region) ∧ mpeg7x:area(region, area) ∧
swrlb:lessThan(area, 80 pixels) ∧ Dense(region) ∧
mpeg7x:solidity(region, solidity) ∧ swrlb:greaterThan(solidity, 0.95)
→ mpeg7:Depicts(region, mesh:Ribosome)

3.7.3 Retrieval results

Table 3.1 shows the results of applying the rules, with the number of relevant regions in the collection (Rel), the number of regions retrieved (Ret) and the number of relevant regions retrieved (RetRel). Precision is defined as RetRel/Ret, recall as RetRel/Rel. Results are shown for retrieval of the Golgi stack (Go), the endoplasmic reticulum (ER), mature granules (MG), mitochondria (Mi), ribosomes (Ri) and the mean of all regions.

The results demonstrate a number of things. Firstly, precision is higher than recall. Given that the aim is to construct a three-dimensional image from a stack of images, this is a good thing: an incorrectly annotated region would cause more problems than a missing region that can be detected by neighbouring slices in the image stack. Secondly, the precision of the rule for ribosomes is 100 %. This can be explained by the way segmentation was performed: a predefined circle with perfect shape and fixed size was used to manually segment the regions depicting ribosomes. This makes it very easy to extract ribosomes based on size and shape. Recall is less than 100 % because

Table 3.1 Rules by domain expert.

	Go	ER	MG	Mi	Ri	Mean
Rel	2486	1463	221	105	1125	
Ret	185	153	1098	346	824	
RetRel	126	37	10	11	824	
Prec. (%)	68.11	24.18	0.91	3.18	100	39.28
Recall (%)	5.07	2.53	4.52	10.48	73.24	19.17

Table 3.2 Ordered rules.

	Go	ER	MG	Mi	Ri	Mean
Rel	2486	1463	221	105	1125	
Ret	487	119	965	183	844	
RetRel	365	36	10	12	844	
Prec. (%)	74.95	30.25	1.04	6.56	100.00	42.56
Recall (%)	14.68	2.46	4.52	11.43	75.02	21.62

regions can overlap, which breaks the perfect circular shape. Finally, mature granules score poorly on both recall and precision. We expect this is due to the large variations in shape that mature granules tend to display.

The order in which rules are executed affects the results, since an annotation given by one rule can not be overwritten by the next rule. Table 3.1 depicts the mean results of all possible orders of execution. To improve the overall results, we sorted the rules in such as way that the most accurate and reliable rules were executed first, while the more general, imprecise rules were executed last. We found a slight improvement in both recall and precision (Table 3.2). The optimal order is Ri, Go, Mi, ER, MG.

One of the benefits of using rules is that they can be easily refined and improved. For example, we can tweak recall and precision by either increasing or decreasing the thresholds in the rules. When *better recall* is more desirable than *high precision*, one can modify the rules so that more regions are included. Using our knowledge of images features, we were able to make slight modifications to the domain expert's rules that improved recall and precision (Table 3.3). We added a rule to extract tubular vesicles. Like ribosomes, they are easily recognisable thanks to the perfect circular shape of the segmentation. By executing this accurate and reliable rule first, we further improved the results.

3.7.4 A comparison to the fuel cell domain

Both recall and precision were significantly higher for fuel cell images than for the pancreatic cell images. The differences between the two domains lie in a number of factors. Firstly, the pancreas domain consists of a higher number of distinct object categories. We did not find, however, that this had a negative effect. Adding one more category (tubular vesicles) even improved the results.

Table 3.3 Rules by the authors.

	Go	ER	MG	Mi	Ri	TV	Mean
Rel	2486	1463	221	105	1125	148	
Ret	4689	48	578	83	855	148	
RetRel	2182	35	207	75	834	148	
Prec.	46.53	75.92	35.81	90.36	97.54	100	73.85
Recall	87.77	2.39	93.67	71.43	74.13	100	71.57

Secondly, the variations in appearance within categories of pancreatic cellular components was greater than in the fuel cell domain. The Golgi stack and the ER, for example, can be seen in many shapes and sizes. This causes the rules to be more complex. Retrieving the Golgi involves utilising the shape and position of neighbouring regions, while the rule to retrieve the ER includes examining and recognising regions in related images. Precision and recall of the Golgi rule are less than that of the results in the fuel cell domain, but still reasonable. The rule to retrieve the ER proved too complex for our system at this stage, due in part to limitations in the segmentation of ER regions and touching ribosomes. The maximum recall we could achieve for ER was 2.4 % which is unacceptable.

Regions in the pancreatic cell images that have similar visual characteristics to the regions in the fuel cell images, namely the mitochondria, ribosomes and tubular vesicles, achieve similar, satisfactory results.

3.8 Discussion

In this chapter we described an interface for annotation of regions in images based on user-formulated semantic inferencing rules. We measured the performance in the domain of pancreatic cell images. The aim of this study was to determine domain characteristics that suit the semantic inferencing and Rules-By-Example approach. One of the significant findings was that knowledge of multimedia and image analysis terms is both a prerequisite and impediment to obtaining good results. We sought to overcome this barrier by providing intuitive graphical tools to formulate the rules and by defining intermediate-level terms for building rules. We still found, however, that the results of applying rules defined by domain experts were significantly less than results of rules defined by the authors. This disparity may be reduced in time as the domain experts' use of the system improves. Other possible solutions include: (1) training of domain experts to familiarise them with multimedia terms and techniques and (2) enabling image analysis experts and domain experts to formulate rules collaboratively.

Our research determined that the segmentation step is important for the quality of the annotations. Ribosomes and tubular vesicles were annotated almost perfectly thanks to their simple shape and the level of segmentation. Clatharin coated vesicles, on the other hand, were a class of cellular components for which the domain experts considered it infeasible to define a rule. The

clatharin coat of these vesicles appears as a cloud on the outside of the segmented region and can therefore not be described by our image analysis techniques. We conclude that our approach depends on strong segmentation, which is the identification of regions in an image that represent conceptual objects. Weak segmentation, on the other hand, is the grouping of pixels into regions, based on homogeneity of low-level visual features (Smeulders et al. 2000).

Our research indicated that the system performs better if the domain's scope is relatively narrow and consists of well-understood concepts that are widely agreed upon, so that subjectivity is minimised. Classes need to display small visual variance and be clearly visibly distinguishable from other classes. Given these characteristics, we expect this approach to be valuable for semantically annotating images in other medical domains, botanical domains (i.e., plant identification), or for analysis of remote surveillance satellite images.

Adapting the system to a new domain proved to be relatively easy. Most of the time and effort went into the process of selecting and extending existing vocabularies to make them suitable for the specific domain. The RDF and OWL versions of MeSH and Mpeg-7 provided us with easily extensible, interoperable, conceptual frameworks. RuleML was not sufficiently expressive for the types of rules required in a complex domain such as pancreatic cell images. The lack of built-in relations and limited integration with RDF/OWL meant that we had to use Python scripts and SQL queries to pre-process the data. This adversely affected the interactivity and responsiveness of the system. We expect that by using SWRL (which combines RuleML and OWL), this problem will be largely overcome.

3.9 Future Work

We see possibilities for improvement in the domain of pancreatic cells by making more use of the three-dimensional nature of the images and the volumetric and spatial relationships between image regions. Considering that one cellular component appears in several image slices, one can complete gaps in annotations of unidentifiable regions by examining the annotations of regions in the surrounding slices. Exploiting the third (depth) dimension could also be valuable in other domains, such as brain scans or mammograms.

In the pancreatic cell domain, the precision of the inferencing rules varied significantly. Applying the rules in order of decreasing precision showed an improvement in the results. We are planning to extend the RBE interface so that domain experts can easily manipulate the order in which rules are executed. We are also investigating possible system upgrades that provide more direct feedback on the effect of the rules. However, this could increase the risk of over-fitting the rules to the data and make them less applicable to other or new image collections in the same domain.

Another option to improve the ease of rule formulation is a fuzzy representation. Concepts like the size of a region or the irregularity of a shape are not easily expressed with precise boundaries and using fuzzy logic might provide a more natural representation. Furthermore, a hybrid approach which incorporates machine-learning techniques to optimise values based on user-defined

combinations and value ranges may also be useful.

If rules are formulated by more than one person, conflicting rules might occur. Future research is needed to determine the best strategy if a region is classified differently by two rules. The number and severity of conflicts might be an indication of the visual complexity of a domain, and therefore an indication of the suitability for approaches such as the one taken in this chapter.

Acknowledgements

Thanks to Brad Marsh and Adam Costin from the Institute of Molecular Biology for their feedback and the images.

4

Assessing User Behaviour in News Video Retrieval

In this chapter we present the results of a study in which we assess information seeking behaviour of people querying a news archive using an interactive content-based image retrieval (CBIR) system. In Chapter 2, we studied user image descriptions in a setting not related to a specific domain or retrieval method. The present study shows the way users formulate queries for news videos using a CBIR system. We investigate whether the size of semantic gap changes with different categories of topics and queries. Correlations between user characteristics, search behaviour, categories of user queries and quality of search results are measured. Based on the results we discuss implications for the design of user interfaces of video retrieval systems.

This chapter, which was co-authored by Giang Nguyen, Dennis Koelma, Guus Schreiber and Marcel Worring, was published in the IEE proceedings on Vision, Image and Signal Processing (Hollink et al. 2005a). An earlier version was published in the proceedings of the International Conference on Image and Video Processing (Hollink et al. 2004c).

4.1 Introduction

In this chapter we study information seeking behaviour of users searching in a collection of broadcast news video. Large collections of broadcast material are maintained at broadcasting stations and at archiving organisations such as ”The Netherlands Institute for Sound and Vision” and the ”Institut National de l’Audiovisuel”. In recent years, these archives have been queried by a broad user group including broadcasters, documentary makers, researchers and students. However, access to broadcast video is still difficult and too often a time consuming process (Hecht et al. 2004).

Many techniques have been developed to automatically index and retrieve multimedia. The TREC Video Retrieval Evaluation (TRECVID)[1] provides test collections and software to evaluate these techniques. Video data and statements of information need (topics) are provided in order to evaluate video-retrieval systems performing various tasks. In this way, the performance of the systems is measured. However, these measures give no indication of how user-behaviour and user-characteristics affect the performance of retrieval systems. Variables like prior search experience and knowledge about the topic can be expected to influence search results. In addition, search behaviour such as the formulation of textual queries and the selection of example images, will influence the results. Due to the recent nature of automatic retrieval systems, little data is

[1] http://www-nlpir.nist.gov/projects/trecvid/

available about user experiences. We argue that knowledge about user-behaviour is one way to improve performance of retrieval systems. Interactive search in particular can benefit from this knowledge, since the user plays such a central role in the process. Studies have been done to measure usability of interactive retrieval systems (e.g. Christel and Moraveji 2004) and effectiveness of different components of these systems (Yang et al. 2004). In this chapter we investigate the still unclear impact of user-behaviour and user-characteristics on the performance of interactive retrieval systems.

We participated in the interactive search task of TRECVID and monitored user behaviour on a state-of-the-art interactive news video retrieval system (Worring et al. 2004). The TRECVID collection consists of 60 hours of video from ABC, CNN and C-SPAN. News data can in theory contain every theme in the world, which complicates the retrieval process. However, this broadness also makes it a valuable test collection, since the results will be applicable to a wide range of collections. Within this broad context, we focus on category search: a user is searching for shots belonging to a certain category rather than for one target shot.

In this study we record data about user characteristics, familiarity of users with topics, queries formulated by users and actions that users take when using the system. In particular, we are interested in which actions lead to the best results. To achieve an optimal search result, a user needs to have a good overview of the contents of the collection. This will give the user an idea of the recall and precision of a search and will aid the user during the search process in deciding whether a continuation of the search is likely to yield new and better results. Therefore, in this study we measure how well users estimate the quality of their search.

We expect that not only user actions, but also the topics influence the results. It is possible that different categories of topics lead to different user actions and differences in the quality of the results. Using a simplification of the framework in Chapter 2, we therefore categorise the 24 topics provided by TRECVID. We compare search behaviour and search results of categories of topics.

In summary, the main questions in the study are:

1. What search actions are performed by users and which actions lead to the best search results?
2. Are users able to estimate the success of their search?
3. What is the influence of topic category on user actions and search results?

4.2 The Interactive Video Retrieval System

4.2.1 Indexing of the video data

Prior to user interaction, the whole collection of video data is indexed in order to provide the user with high-level entry points into the data set. Firstly, we derive high-level textual concepts from the automatic speech recognition (ASR) result (Gauvain et al. 2002) using Latent Semantic Indexing (LSI) (Deerwester et al. 1990). To that end, we construct a vector space by taking all words found in the ASR results of all videos in the collection. We then perform stopword removal

using the SMART's English stoplist. This results in a 18,117 dimensional vector space. Using LSI the vector space is reduced to 400 dimensions. Thus, we decompose the information space into a small set of broad concepts, where the selection of one word from the concept reveals the complete set of associated words also.

Secondly, we use 17 high-level concept detectors developed by Carnegie Mellon University (CMU) for the TRECVID (Hauptmann et al. 2003), ranging from generic ones like outdoors to more specific ones like physical violence. The quality of the detectors ranges from poor to good.

In addition, for all keyframes in the data set we perform low-level indexing by computing the colour histograms using 32 bins for each channel. To structure these low-level visual descriptions, the whole data set is clustered using k-means clustering with random initialisation. The k in the algorithm is set to 143 as this is the number of images our display will show to the user.

4.2.2 User interaction with the system

User interaction with the system consists of two steps: (1) filtering of the complete data set into a smaller 'active set' and (2) browsing through the active set. A user enters the system with an information need. In TRECVID, statements of information need are statements like 'find shots of an airplane taking off', or 'find shots of the Sphinx'. A typical session on the system starts with a user entering a textual query (Figure 4.1). The user then chooses between 'exact search' (without LSI) or 'concept search' (with LSI). By default the system is set to 'concept search' (Figure 4.1). In addition, the user can indicate the desired presence or absence of each of the 17 high-level concepts. Users can combine the two query mechanisms using an 'and' function (but this usually leads to very small sets and low recall) or an 'or' function, where the ranked result is an alternation between the results obtained for the selected query specification mechanisms. The default value is 'or'. The two mechanisms together produce a ranked list of shots, the active set, that is used in the subsequent browsing step. We restrict the active set to contain a maximum of 2000 shots, leading to approximately 4000 keyframes.

In the browsing step, keyframes from the active set are displayed to the user. Browsing requires a visualisation mechanism that on the one hand provides an overview of the data set, while showing sufficient detail on the other. Furthermore, the visualisation should give the user insight into the structure of the data set. The system supports the user with an array-based (Figure 4.2) and a similarity-based (Figure 4.3) visualisation. When the user points to a thumbnail of a keyframe, a full size image and text associated with the shot are shown on the right side of the screen. Sequential keyframes in the video from which a keyframe is selected, are presented at the bottom of the screen (Figure 4.3).

The user can now select relevant example keyframes from within the active set. When the user has selected a set of examples, he or she can click the 'feedback' button in order to obtain a ranked result list of images from the active set. Ranking of the active set is based on query-by-example (QBE) where similarity of two keyframes is defined by the Euclidean distance of the two colour histograms. In the result the closest matches with the example images are computed, where the

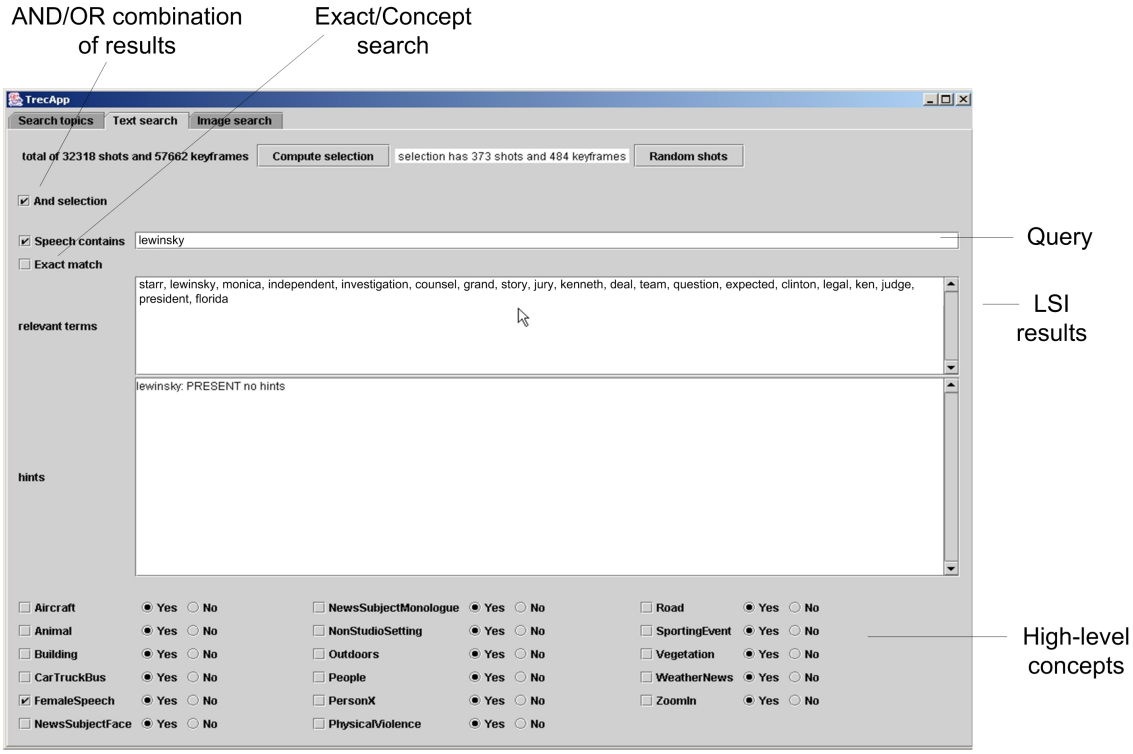

Figure 4.1 Screen shot of the GUI used for query entering.

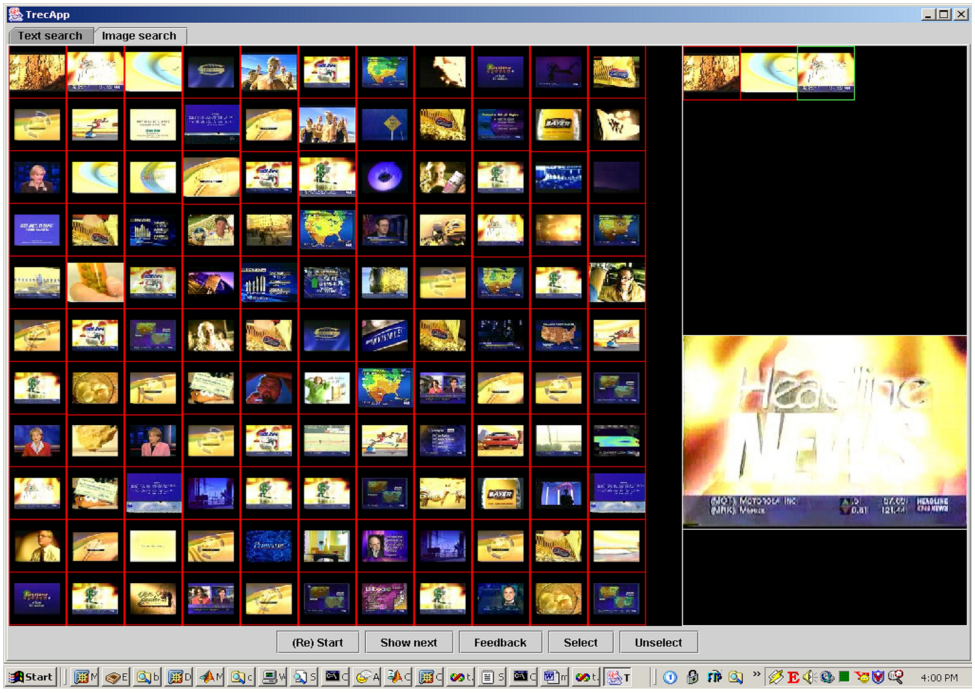

Figure 4.2 Screen shot of the GUI used for browsing with array-based visualisation.

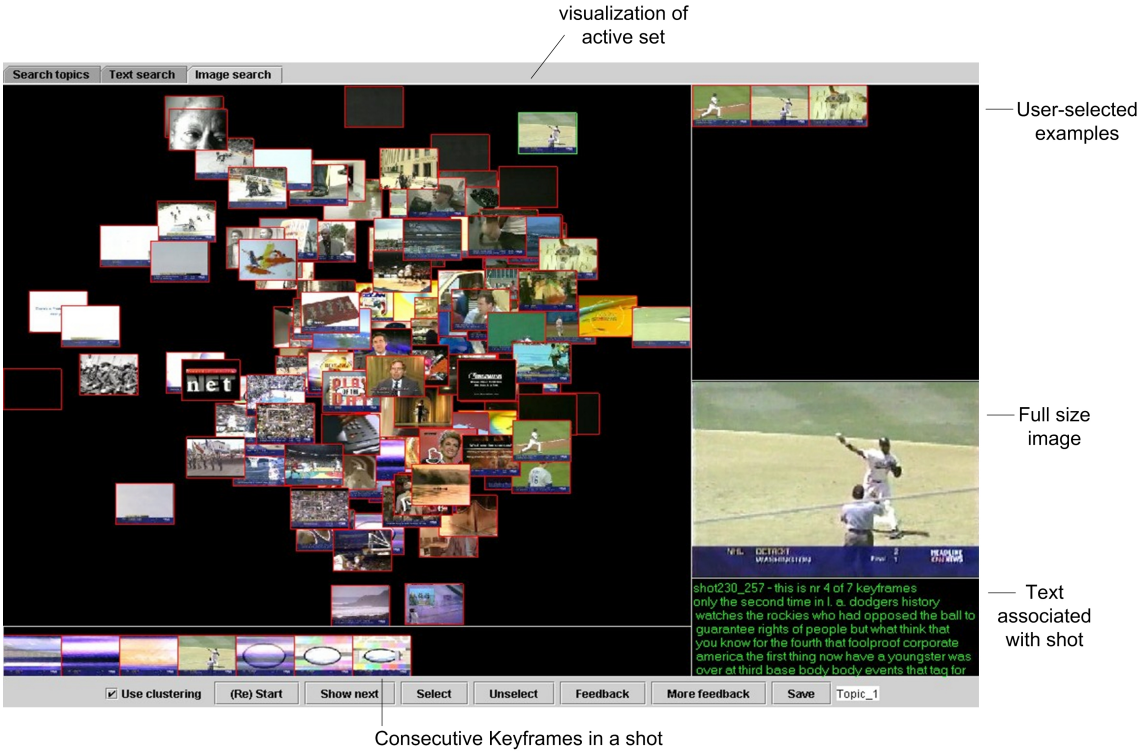

Figure 4.3 Screen shot of the GUI used for browsing with similarity-based visualisation.

system alternates between the different examples selected. The user-selected example images are placed in the highest ranks of the result list. To allow for easy comparison between systems, we follow the TRECVID custom by always letting the result of a search consists of a ranked list of 1000 items.

If the resulting ranked list of keyframes is not satisfactory, the user can decide to go back to the filtering stage and change the query, or to continue browsing for relevant examples and perform a new ranking. The re-ranked set is again visualised as in Figure 4.2 or 4.3, enabling the discovery of new relevant keyframes. The process of querying, browsing and (re)ranking continues until the user is satisfied and saves the results.

4.3 Related Work: A Comparison to Other Systems

User actions on a system always take place within the bounds of the user interface of the system. Comparing the user interface of the current system to similar systems gives an indication of how much of the user-behaviour is specific to the current system and how much is likely to be typical to interactive video retrieval systems in general. In the TRECVID 2003 conference, 12 systems participated in the interactive search task (Smeaton et al. 2003). We compare the features of our system to 5 well performing systems from IBM (Amir et al. 2003), Carnegie Mellon University

(Hauptmann et al. 2003), Dublin City University (Browne et al. 2003), Imperial College London (Heesch et al. 2003) and the University of Oulu (Rautiainen et al. 2003). Some of the systems come in multiple forms (e.g a text only system and a combined text/image system). In these cases we look at the most extensive variant. We do not seek to do justice to all design decisions that have been made in these systems. Instead, we try to give a short overview of the major functionalities of the user interfaces. We will not go into the underlying techniques that are hidden from the user, even though these techniques are no doubt of major importance for differences in performance.

All systems provide the user with a field in which free text queries can be typed and all systems use ASR results to process textual queries. Also, all systems have a mechanism to let users select positive example images while browsing. Positive examples can be added to the query and will appear high in the result list. The system from Dublin adds not only the keyframe, but also the associated text to the query when a keyframe is selected as a positive example. The IBM system allows users to select negative examples (Amir et al. 2003). Two systems (Hauptmann et al. 2003, Rautiainen et al. 2003) use the high-level concepts from the TRECVID feature task (e.g. Car/Truck/Bus, Female Speech, Outdoor) to filter the data set in a similar fashion to our system. None of the systems uses LSI to extend textual queries.

There are different ways to combine the components of a query (text, example images, high-level concepts). Two systems let users adjust the weights of textual and image queries (Amir et al. 2003, Browne et al. 2003). The system from Oulu lets users switch text-based, image-based and high-level-concept-based search on or off. In the London system users can perform relevance feedback by moving keyframes around the screen.

All systems offer the user a way to visually inspect the result-set as a list of keyframes ranked in the order of similarity with the query (e.g. Figure 4.2). The system from London (Heesch et al. 2003) has an alternative view where the layout of keyframes on the screen visualises similarity of keyframes with the query. This is similar to our system, that visualises the similarity between keyframes (Fig 4.3). Most systems use the temporal aspect of video by showing sequential keyframes of a shot within the ranked result list (Amir et al. 2003, Browne et al. 2003, Hauptmann et al. 2003, Rautiainen et al. 2003). The London system (Heesch et al. 2003) shows sequential keyframes in a separate window triggered by a user selecting a keyframe, similar to our series of consecutive keyframes on the bottom of Figure 4.3. Most systems provide users with a way to inspect a single keyframe in a larger window (Amir et al. 2003, Browne et al. 2003, Hauptmann et al. 2003, Heesch et al. 2003) and some let the user play the video (Amir et al. 2003, Browne et al. 2003, Hauptmann et al. 2003).

We can conclude that there is a considerable overlap in functionalities. The querying and browsing interfaces show similarities across all systems. The main points that are specific to our system are the way of combining different retrieval mechanisms and the use of LSI to facilitate concept search.

4.4 Methods

During the study, 21 groups of subjects (18 pairs and 3 individuals) searched the system for 12 topics per group. Prior to the study, subjects had received a three-hour training on the system. The data were analysed on the level of individual searches. A search is defined as the process of one subject group going through the three interactive stages of the system for one topic. After exclusion of searches that were not finished or contained too much missing data, 242 searches remained. To prevent sequential scanning of all shots in the collection, the time to complete one search was limited to 15 minutes.

Four types of data were gathered: average precision of a search, data about the interaction during a search, user estimation of the quality of a search and the category of topics and queries.

Average precision

Average precision (AP) was used as the measure of quality of the results of a search. AP is the average of the precision value obtained after each relevant camera shot is encountered in the ranked result list (NIST 2005). The value of AP lies between 0 and 1 and favours highly ranked relevant camera shots. Let $L^i = \{l^1, l^2, ..., l^i\}$ be a ranked version of the answer set A. At any given index i let $|R \cap L^i|$ be the number of relevant camera shots in the top i of L, where $|R|$ is the total number of relevant camera shots. Then AP is defined as:

$$\text{AP} = \frac{1}{|R|} \sum_{i=1}^{|A|} \frac{|R \cap L^i|}{i} \lambda(l^i)$$

where $\lambda(l^i)$ = 1 if $l^i \in R$ and 0 otherwise. Note that AP is a quality measure for *one search* and not the mean quality of a group of searches. The iterative process of querying, browsing and ranking causes the AP of the result-set to rise and fall during the search. Therefore, we recorded not only the AP at the end of the search but also the maximum AP during the search.

AP of each search was computed with a ground truth provided by TRECVID. Shots that were not in the ground truth were judged as being “not relevant”. This is not always correct, since the ground truth contains only shots that were retrieved by the TRECVID participants. We do not consider this a problem since all searches in our study suffer from the same disadvantage.

Search data

In order to answer the first research question, logs of user interactions with the system were made containing the following data about each search:

1. duration of the search
2. number of textual queries
3. high-level concepts that were used
4. number and type of images selected

5. whether AND or OR search was used
6. whether exact (without LSI) or concept (with LSI) search was used

These data were examined at two points in time: at the end of the search and at the point at which maximum average precision was reached.

User estimation

To answer the second research question, a questionnaire was developed to measure user estimation of the success of a search. Four questions were answered after each search:

1. Was it easy to get started on this search?
2. Was it easy to do the search on this topic?
3. Do you expect that the results of this search contain a lot of non-relevant items (low precision)?
4. Are you satisfied with your search results?

   In addition, subjects answered the following question after each search:
5. Are you familiar with this topic?

All questions were answered on a five-point scale (1=not at all, 5=extremely).

Categories of topic descriptions and textual queries

The 24 topics provided by TRECVID and the textual queries formulated by the subjects were categorised using a framework that was designed for a previous study (Chapter 2). The framework combines different methods (e.g. Armitage and Enser 1997, Jörgensen 1998) to categorise image descriptions into various levels and classes. For the present study we used only those distinctions that we considered relevant to the list of topics: ‘general’ vs. ‘specific’ and ‘static’ vs. ‘dynamic’. Other distinctions, such as ‘object’ vs. ‘scene’, were not appropriate for the topic list since most topics contained descriptions of both objects and scenes. A summary of categorised topics is provided in Table 4.1.

4.5 Subjects

All subjects were students in Information Science who enrolled in the course Multimedia Retrieval at the University of Amsterdam. The number of years of enrollment at the university was between 1 and 8 (mean = 3.5). Two subjects were female, 37 were male. Ages were between 20 and 40 (mean = 23.4).

In order to control to what extent prior search experience might interfere with the effect of search actions on the results, we asked the subjects to fill in a questionnaire that contained questions about frequency of use and experience with information retrieval systems in general and,

Table 4.1 Summary of topics, categorised into general and specific and into dynamic and static. See http://www.cs.vu.nl/˜laurah/trec/topics.html for topic details.

Class	General	Specific
Static	01: aerial view of buildings and roads	09: the mercedes logo
	06: helicopter in flight or on ground	25: the white house
	10: one or more tanks	07: tomb of the unknown soldier
	13: flames	17: the sphinx
	14: snow-covered mountains and sky	24: Pope John Paul II
	16: road(s) with lots of vehicles	04: Yassar Arafat
	18: a crowd in an urban environment	20: graphic of Dow Jones
	22: cup of coffee	15: Osama bin Laden
	23: cats	19: Mark Souder
Dynamic	05: airplane taking off	02: basketball passing down a hoop
	12: locomotive approaching you	03: view from behind catcher while pitcher is throwing the ball
	08: rocket taking off	
	11: person diving into water	

more specifically, with multimedia retrieval systems. It appeared that all subjects searched for information at least once a week and 92 % had been searching for two years or more. All subjects searched for multimedia at least once a year and 65 % did this once a week or more. 88 % of the subjects had been searching for multimedia for at least two years. We did not find any evidence of a correlation between prior search experience and actions, nor between prior search experience and search results.

After the study, all subjects filled in a short questionnaire containing questions about the user's opinion of the system and the similarity between this type of search and the searches that they were used to perform. All but three subjects indicated that the system was not at all similar to what they were used to. All subjects disagreed with or were neutral to the statement that the topics were similar to topics they typically search for. The lack of influence of search experience can in part be explained from the fact that the system was different from search systems that the subjects were used to. 78 % feel that the system is easy to use.

The subjects indicated a high familiarity with the topics[2]. Spearman's correlation test indicated a relationship between familiarity and AP only within topics 10 and 13. We do not consider this enough evidence that there is in fact a relationship.

[2] An exception was topic 19 "Find shots of congressman Mark Souder", with whom none of the subjects was familiar

Table 4.2 User actions in the system at the moment of maximum AP and at the end of the search.

		Max				End			
User Action	N	Min.	Max.	Mean	St.D.	Min.	Max.	Mean	St.D.
Time to finish topic (sec.)	242	0	852	345	195	6	899	477	203
No. of query (re)formulations	220	.	.	.	.	1	25	7.51	5.31
No. of high-level concepts used	240	0	5	0.50	0.84	0	17	0.59	1.39
No. of images selected	242	0	30	8.47	7.01	0	30	9.07	7.06
AND or OR search	240	AND:75 OR:165				AND:82 OR:158			
Exact or Concept search	240	exact:69 concept:166				exact:62 concept:176			

4.6 Results

4.6.1 Search data

The first research question was ‘what search actions are performed by users and which actions lead to the best result?’ In Table 4.2 descriptives are presented of the six data types that were recorded in the user logs. It shows that a search took approximately 8 minutes; a mean of 7.5 different textual queries were formulated during a search; a mean of 9 images were selected per search; high-level concepts were hardly used; or-search was used more than and-search; concept search (with LSI) was used in most cases. Out of the six variables that were measured, ‘duration of the search’, ‘number of textual queries’, ‘and/or search’ and ‘exact/concept search’ did not affect the AP of the result. The number of high-level concepts that were used had a negative influence on the AP, while the number of example images that was selected had a large positive influence. In the next subsections we will discuss the number of textual queries, the use of high-level concepts and the selection of example images.

Query (re)formulation.

In total, the subjects formulated 2141 textual queries. This brings the mean number of textual queries per search to more than seven. Going back and forth between the different stages of the retrieval process and reformulation of the query, is apparently an important part of user behaviour. This corresponds to the findings of Goodrum et al. (2003), who examined image searching behaviour of users on the web. She found that query reformulation was one of the frequently occurring patterns of search tactics. The number of queries did not affect the AP of the result.

High-level concepts.

The number of high-level concepts that was used in a search had a negative influence on the result ($r = -0.17$, $p < 0.01$). This is depicted in Figure 4.4. The number of uses per high-level concept was too low to draw conclusions about the quality of individual concepts. We can conclude,

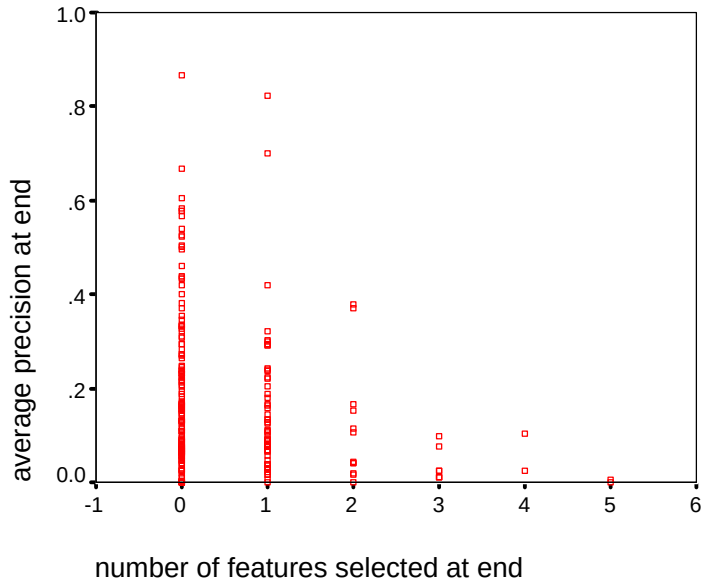

Figure 4.4 Scatter plot of number of selected concepts and AP at the end of the search. One case with 17 concepts and AP of 0.027 is left out of the plot.

however, that selection of more than one concept leads to low average precision. To give an indication of how the concepts were used by the subjects, Table 4.3 shows the frequency of use of the concepts and the mean AP of searches using the concepts. Only searches in which a single concept was used are included. Improving the quality of the concept detectors might lead to more use of the concepts and better results when concepts are combined. Snoek et al. (2004) showed a great improvement in detector performance.

Example images.

The number of selected images was the most important variable to explain the result of a search (Pearson's correlation coefficient r = 0.37, $p < 0.01$). This can be explained from the fact that each correctly selected image adds at least one relevant image to the result-set. The contribution of the ranking to the result was small; change in AP caused by the ranking step had a mean of 0.001 and a standard deviation of 0.032. The mean average precision at the end of a search was 0.16. The number of selected images was not correlated to the time to finish a topic, to the number of high-level concepts used, or to the type of search.

4.6.2 User prediction of search quality.

User estimation.

We collected opinions and expectations of users on each search. All questions measure an aspect of the user's estimation. For each question a high score represents a positive estimation, while a

Table 4.3 High-level concepts: number of times a concept was used, mean average precision of searches using this concepts and standard deviation of the average precision.

Concept	N	AP	St.d.	Concept	N	AP	St.d.
Aircraft	5	0.09	0.05	People	3	0.13	0.15
Animal	5	0.17	0.06	PersonX	7	0.14	0.16
Building	2	0.30	0.00	PhysicalViolence	0	.	.
CarTruckBus	4	0.11	0.03	Road	3	0.06	0.04
FemaleSpeech	0	.	.	SportingEvent	9	0.08	0.03
NewsSubjectFace	1	0.24	.	Vegetation	1	0.13	.
NewsSubjectMonologue	1	0.70	.	WeatherNews	0	.	.
NonStudioSetting	4	0.15	0.13	ZoomIn	1	0.08	.
Outdoors	15	0.17	0.20				

Table 4.4 Results of the Principal Component Analysis.

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cum. %	Total	% of Variance	Cum. %
1	2.78	70	70	3	70	70
2	.67	17	86			
3	.34	9	70	70	3	70
4	.20	5	17	86		

Table 4.5 Loading of questions on component 1.

Questionnaire item	Loading on Component 1
easy to start search	0.87
easy to do search	0.91
satisfied with search	0.87
expect high precision	0.66

low score represents a negative estimation. Mutual dependencies between the questions complicate conclusions on the correlation between each question and the measured average precision of a search. Therefore, we combined the scores on the 4 questions into one variable using Principal Component Analysis. The new variable that is thus created represents the combined user estimation of a search. This variable explains 70 % of the variance between the cases (Table 4.4). Table 4.5 shows the loading of each question on the first principal component. Pearson's correlation test showed a relationship between combined user estimation and actually measured average precision. (r = 0.30, $p < 0.01$). This suggests that users are indeed able to estimate the success of their search.

Time between maximal AP and the end of the search.

Another measure of user estimation of a search is the difference between the point where maximum precision was reached and the point where the user stopped searching. As mentioned in Section 4.6.1, the mean time to finish a search was 477 seconds, while the mean time to reach maximum average precision was 345 seconds. The mean difference between the two points in time was 128 seconds (min = 0; max = 704, sd = 142). This means that subjects typically continued their search for 128 seconds (more than two minutes) after the optimal result was achieved. This suggests that even though subjects were able to estimate the overall success of a search, they did not know when the best results were achieved within a search. Not knowing when to stop searching is a general problem of category search.

A correlation between combined user estimation and time-after-maximum-result shows that the extra time was largest in searches that got a low estimation ($r = -0.426$, $p < 0.01$). The extra 2 minutes did not do much damage to the precision. The mean average precision of the end result of a search was 0.16, while the mean maximum average precision of a search was 0.18. The mean difference between the two was 0.017 (min = 0; max = 0.48; sd = 0.043).

4.6.3 Topic and query category

Topic type

Table 4.6 shows that 'specific' topics were better retrieved than 'general' topics. The results of 'static' topics were better than the results of 'dynamic' topics, which can be explained by the fact that our system treats the video data in terms of keyframes, which are still images. A two-way analysis of variance showed significant effects of both factors specific/general (F = 22.8, $p < 0.01$) and static/dynamic (F = 22.2, $p < 0.01$), as well as a significant interaction effect (F = 34.8, $p < 0.01$). However, within the general topics, an effect of static/dynamic topics on AP was not found (t = -1.3, df = 148, p = 0.18). Likewise, within the dynamic topics, an effect of general/specific topics on AP was not found (t = 1.4, df = 68, p = 0.18). We conclude that the results of *specific static* topics are better than the results of the other categories of topics. There is a strong correlation between topic-category and query-category. However, the correlation between topic-category and

Table 4.6 Mean AP of topics types.

Mean AP	Static	Dynamic	Total
General	0.08	0.10	0.09
Specific	0.28	0.08	0.24
Total	0.19	0.10	0.16

Table 4.7 Query categories: absolute numbers at max. AP and at the end of the search and table percentages.

Query Type	From Topic	Not From Topic	Total
general	93 + 78 (37%)	68 + 79 (32%)	161 + 157 (69%)
specific	42 + 49 (20%)	28+ 25 (11%)	70 + 74 (31%)
Total	135 + 127 (57%)	96 + 104 (43%)	231 + 231 (100%)

AP is valid regardless of the query-category used.

The change in AP caused by the ranking step was positive for general topics (mean change = 0.005), while negative for specific topics (mean change = -0.004). For general topics we found a correlation between *change in AP* and *AP at the end of the search* (r = 0.265, $p < 0.01$), which was absent for specific topics.

Query type

The length of textual queries varied from 1 to 22 words. To avoid domination of the analysis by long queries, we used only the first word of every query to determine query type. During a search, users may formulate multiple textual queries. We analysed the last query before the moment of maximum AP and the last query of the search (Table 4.7). 69 % of the queries formulated by the subjects was ‘general’, 31 % was ‘specific’; 93 % was ‘static’ and 7 % was ‘dynamic’. Considering the low number of dynamic queries (only 16), we limit further analysis to the distinction between ‘general’ and ‘specific’ queries.

A total of 57 % of the query words were copied directly from the topic descriptions. In the copied queries, the share of specific terms was higher than in queries that were not taken from a topic. An analysis of variance showed that ‘specific’ queries led to better results than ‘general’ queries (F = 30.1, $p < 0.01$). This is still true for the ‘general’ topics: some subjects used specific queries to search for general topics (e.g query for Micheal Jordan, when looking for shots of basketball games) and that strategy worked very well. We did not find any evidence that other user actions were different for different topic categories.

4.7 Discussion

This study was concerned with the question how users search for news video in an interactive video retrieval system and what factors influence the quality of their search results. The results exposed the two aspects of user search behaviour that have the largest impact on the quality of the search results: (1) a high number of selected example images increases the quality of the results, and (2) 'specific' queries give better results than 'general' queries.

The study has been carried out in one domain (broadcast news) using one retrieval system. Future research is needed to see whether the results can be extended to other domains and systems. Our expectation is that the broad domain of news will capture a lot of the difficulties in other domains. The specific structure of news videos – short stories about one topic – was not used by the retrieval system. A comparison of the user interface of the present system to user interfaces of other systems showed a considerable overlap in functionalities. This strengthens our belief that the conclusions and recommendations that we present in this section extend beyond this one system. In the next subsections we will discuss the results of the study and comment on implications for the design of user interfaces of future systems.

4.7.1 Textual queries

The contribution of the ranking step to the average precision was extremely small. From this we can conclude that text is a central feature in news video retrieval. This might change over time as performance of CBIR improves. The importance of text for video retrieval has not gone unnoticed in the TRECVID conferences and was pointed out by Hauptmann (2004), amongst others. Eakins (2004) pointed out that users of image retrieval systems rate text entry interfaces higher than CBIR techniques such as QBE. This point should be taken into account when designing user interfaces of retrieval systems. Supporting text search could, for example, be done by highlighting the words in the retrieved shot that match the user's query.

4.7.2 Topic type

'Specific' topics are better retrieved than 'general' topics. This is in accordance to the average TRECVID results (Smeaton et al. 2003). In our study, the ranking step had a small but positive effect on general topics, while it had a small negative effect on specific topics. This suggests that a different strategy is optimal for different topic types: emphasis should be more on text for specific topics, while it can be on both text and low-level visual features for general topics. Yang et al. (2004) also found that text is especially important for specific topics, while text and QBE are both of importance to generic topics. Allowing the user to adjust the weights of the two retrieval mechanisms, as is done by Amir et al. (2003) and Browne et al. (2003), is a good solution for expert users, but not for beginners as it requires the user to know about the strengths and weaknesses of the retrieval mechanisms. Future retrieval systems could benefit from a (automatic or manual) classification of the topics, in order to adapt the retrieval strategy.

4.7.3 Browsing

The results show that from the recorded user actions, number of selected images is by far the most important variable to explain the success of a search. We conclude that the main contribution of content-based image retrieval to the retrieval process is visualisation of the data set, which gives the user the opportunity to manually select relevant keyframes. The visualisation of the data set also gives the user an overview of the data and thus an indication of the success of the search. The results of the study show that users can estimate success quite well, but do not know when the optimal result is reached within a search. Effective visualisation of the data set and improved facilities for browsing are therefore essential in future retrieval systems. In Nguyen and Worring (2004) an improved version is described of the current similarity-based visualisation of our system (Figure 4.3), that gives a better overview of the data set. Due to the recent nature of automatic retrieval systems, not much is known about the effectiveness of browsing interfaces for video. Van Houten et al. (2004) presented new ideas for a browsing interface in. It would also be interesting to compare the results of an interactive video retrieval system to sequential scanning of shots in the data set for a fixed amount of time.

4.7.4 Background knowledge

Prior experience with searching did not affect the quality of the search results. A possible effect could have been obscured by the three-hour training before the study, or by the fact that most subjects worked in pairs. However, Fang and Salvendy (2000) reported similar results. In their study, prior experience with search tools did not affect the success of searches on the web. Likewise, familiarity with the topic did not affect the quality of the search results. This seems to indicate that background knowledge of the searcher about the topic can not be used adequately in the search process of current retrieval systems. Some attempts to include background knowledge into the process of multimedia retrieval have been made (e.g. Hyvönen et al. 2004b, Jaimes et al. 2003), but inclusion of background knowledge in interactive video retrieval systems is still in an early stage. We believe that text-based search could benefit from structured background knowledge in the form of ontologies or thesauri. This could, for example, be done by linking words in the query to concepts in an ontology, so that synonyms, related terms, broader and narrower terms can de found. In a similar fashion, we expect that search using detection of high-level concepts could benefit from ontologies; by linking each detectable concept to a concept in the ontology, mutual relationships between the concepts can be exploited.

5

Semantic Annotation of Image Collections

This chapter studies the benefits of using structured background knowledge for annotation and search of paintings. The vocabularies that are available for this domain are discussed, together with a metadata schema. The schema is based on findings in Chapter 2. By means of use cases we explore how links between and within vocabularies can be used to support users in the annotation and search process.

This chapter was presented at the Workshop on Knowledge Markup and Semantic Annotation at K-Cap (Hollink et al. 2003), and was co-authored by Guus Schreiber, Jan Wielemaker and Bob Wielinga.

5.1 Introduction

In this chapter we show how ontologies can be used to support annotation and search in image collections. The need for disclosure of image collections is especially large in the cultural heritage sector. More and more cultural heritage institutions make their collections available online in a searchable way. To facilitate access for scholars and visitors, detailed descriptions of the items in their collection are needed.

Cultural heritage institutions have traditionally made use of controlled vocabularies. These vocabularies aid users in finding the right terms, and increase consistency in the use of annotation and search terms. Several large vocabularies have been built for the cultural heritage domain, such as the Art and Architecture Thesaurus (AAT), the Union List of Artist Names (ULAN) and Iconclass. However, many institutions still use their own in-house indexing scheme that typically supports a keyword-type search (Graham 1999).

In the present chapter, we explore how structured vocabularies can be used to improve the annotation and search process in the cultural heritage domain of paintings. We employ multiple large, existing ontologies (WordNet, AAT, ULAN and Iconclass), represented in RDF Schema (Brickley and Guha 2004), to form one controlled vocabulary. Using existing, widely-known vocabularies represented in a standardised language makes the interpretation of annotations and queries easier for anyone who is aware of these vocabularies. The use of such a large, heterogeneous vocabulary raises questions as to how users can be aided in finding the right terms, and how homonymous terms can be disambiguated. In addition, we explore how relations between concepts in the ontologies can be used in annotation and search.

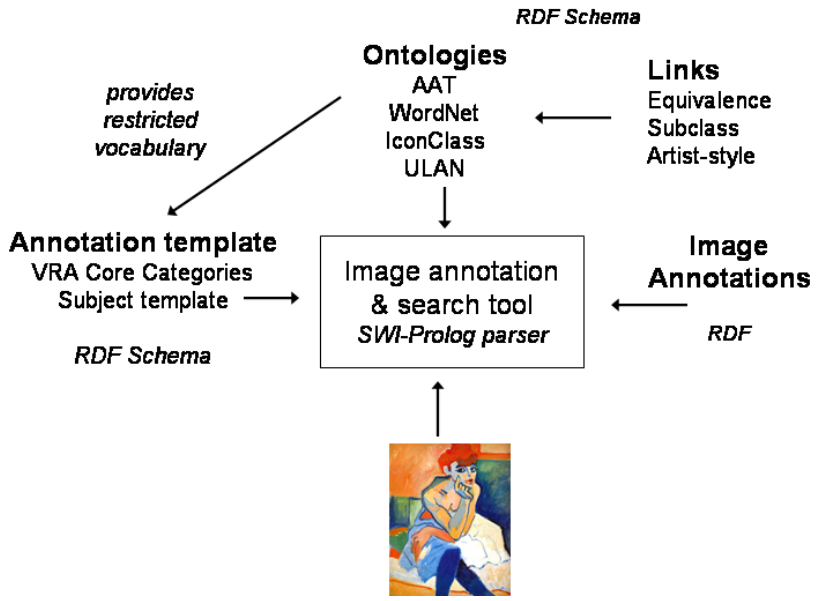

Figure 5.1 Overview of the approach of the MIA annotation and search demonstrator.

We discuss a semantic annotation and search system, called the Mia demonstrator (MIA 2002). Figure 5.1 shows the general architecture of the system. The four ontologies are read into the tool with help of the SWI-Prolog RDF parser (Wielemaker 2000). The tool generates a user interface for annotation and search based on an RDF Schema specification. The tool supports loading images and image collections, creating annotations, storing annotations in RDF and two types of image search facilities: simple search where one search-slot matches all annotation-slots and advanced search, where each annotation-slot can be queried separately. A small experiment concerning the usability of the annotation tool is reported in Schreiber et al. (2002).
In the present work, we use images from the Artchive (Harden 2006), a collection of paintings that are accompanied by short semi-structured textual annotations. The Chagall painting in Figure 5.2, for example, is accompanied by:

Chagall, Marc
Birthday
1915
Oil on cardboard
31 3/4 x 39 1/4 in.
The Museum of Modern Art, New York

Using custom made parsers, we were able to extract the following annotations directly from the text: title, creator, date, materials(medium), materials(support), measurements and current location of the painting. This example shows that valuable annotation time can be gained from

Figure 5.2 ‘Birthday’ by Chagall.

simply reusing existing annotations.

The chapter is structured as follows: Section 5.2 contains related work. In Section 5.3 we discuss the ontologies and a metadata schema that can be used in conjunction with the ontologies. The schema facilitates both art-historic annotations, such as creator and style, and content annotations, describing what is depicted in an image. The annotation and search process is discussed in Section 5.5 in the form of an application scenario. We will show how ontologies are used to find the right term, to disambiguate terms and how links between concepts can be used in annotation and search. Finally, Section 5.6 provides a discussion on the approach taken. The current work is a continuation of work in Schreiber et al. (2001) about semantic annotation and search of a collection of photographs of apes. In the earlier study the emphasis was mainly on the content of the image. For art images both the content and the art-historic features, such as artist and style, are important. This requires the use of additional ontologies (AAT, ULAN) and poses research questions with respect to the links between ontologies (see Section 5.4).

5.2 Related Work

5.2.1 Metadata standards

The Dublin Core Metadata Initiative (DCMI) aims to provide a metadata standard that enables discovery of resources across people, systems and domains. They provide a set of 15 descriptors, called the metadata element set, to describe information resources: title, creator, subject, description, publisher, contributor, data, type, format, identifier, source, language, relation, coverage and rights (Dublin Core 2006). The element set is a widely used ISO standard (ISO 15836:2003(E)). In Clayphan et al. (2005), for example, it is used to support cross-searching of multiple databases. The appeal of Dublin Core is that descriptions are easy to create and understand and also easy to extend.

An extension to the Dublin Core elements set has been made by the Visual Resources Association (VRA 2002). The VRA Core Categories are defined as a specialisation of the Dublin Core set of metadata elements, tailored to the needs of visual resources. Qualifiers are defined for some elements in order to identify data values more precisely. Examples are measurements.dimensions and measurements.resolution for the measurements element and material.medium or material.support for the material element. The VRA Core Categories follow the 'dumb-down' principle, which means that a tool not aware of VRA can interpret the VRA data elements as Dublin Core data elements.

5.2.2 Annotation systems

The architecture shown in Figure 5.1 is in the same spirit as the one described by Lafon and Bos (2002). The main difference lies in the fact that we place more emphasis on the nature of the ontologies. Koivunen and Swick (2003) discuss an architecture for semantic annotation, but mainly from the perspective of the shared collaborations. CREAM (Handschuh and Staab 2003b) also provides an architecture for semantic annotation including both manual and semi-automatic techniques. The present chapter differs from the latter two approaches through its focus on images (which creates special problems, such as annotating the image content) and the practical work on integrating multiple existing ontologies. The work of Hyvönen et al. (2003) combines ontology-based image retrieval with view-based and topic-based retrieval and is probably closest to the present chapter. As of yet, they have not reported many details on the ontologies being used.

5.3 Ontologies and Metadata Schema

A structured annotation requires two things: (1) ontologies or other structured vocabularies of terms or concepts to describe the annotated item and (2) a schema which represents agreement on how to use the vocabulary. A metadata schema consists of slots or properties that can be filled in with values from the ontologies. The Dublin Core element set and the VRA core categories are both examples of metadata schemas.

5.3.1 Ontologies

Four large ontologies are used to fill in the slots of the schema: WordNet, AAT, ULAN and Iconclass.

WordNet WordNet is a general lexical database in which nouns, verbs, adjectives and adverbs are organised into synonym sets (synsets), each representing one underlying lexical concept (Fellbaum 1998). WordNet concepts (synsets) are typically used to describe the content of the image. In this chapter we used WordNet version 1.5, limited to hyponym relations.

AAT The Art and Architecture Thesaurus (AAT) is a large thesaurus containing some 125,000 terms relevant for the art domain. The terms are organised in seven facets, such as the 'Styles

Table 5.1 Number of RDF triples in the four ontologies.

Source	Triples
WordNet 1.5 (limited to hyponym relations)	280.558
Art and Architecture Thesaurus	179.410
Iconclass (partial)	15.452
ULAN (limited to painters)	100.607
Total	576.027

and Periods' facet, the 'Activities' facet and the 'Materials' facet (The Getty Foundation, 2006a). Within the facets, terms are organised by grouping terms, such as <furnishings by form or function> and <furnishings by location or context>.

ULAN The Union list of Artist Names (ULAN) contains information about approximately 220,000 artists. The information includes name variants and some biographical information such as dates, locations and artist types (The Getty Foundation, 2006d). A subset of 30,000 artists, representing painters, is incorporated in the tool.

Iconclass Iconclass is an iconographic classification system, providing a hierarchically organised set of concepts for describing the content of visual resources[1] (Berg, van den 1995). We used a subset of Iconclass.

AAT, WordNet, Iconclass and ULAN were all translated into the RDF Schema notation. For example, WordNet was represented in the following fashion:

- WordNet synsets were represented as RDFS classes;
- word forms of synsets were represented as RDFS labels of the corresponding class;
- hyponym relations were represented as RDFS subclass relations;
- glossary entries of concepts were represented as RDFS comments.

In Wielemaker et al. (2003) the use of WordNet 1.6 as represented in RDF by Melnik and Decker is discussed.[2] In Wielinga et al. (2001) one can find a discussion on issues arising when representing AAT in RDF Schema.

Table 5.1 shows the number of RDF triples in the tool for each of the ontologies. The infrastructure of our current tool can handle this set of 576,000 triples efficiently, but it is expected to break down when the triple base becomes significantly larger. Based on the experiences in this work, a revised version of the infrastructure has been constructed that should be able to handle up to 40,000,000 triples (Wielemaker et al. 2003).

[1] See also http://www.iconclass.nl/

[2] See http://www.semanticweb.org/library/#wordnet for the RDF version of WordNet by Melnik and Decker.

5.3.2 Metadata schema

For our detailed annotation and query purposes, using the ontologies in Section 5.3.1 is not enough. This will disambiguate Paris (France) from Paris (Texas), but it will not disambiguate a painting currently being displayed in Paris (e.g in the Musee d'Orsay) and a painting depicting Paris (e.g. 'Boulevard Montmartre' by Camille Pissarro). Using a schema that provides properties to link ontology concepts to the painting will overcome this problem. In this section, we will describe the metadata schema that we made for paintings and the decisions that lead to its final form.

In Chapter 2, we divided descriptions of images into three levels: (1) low-level perceptual information, such as the colours and shapes that are visible in an image, (2) conceptual information about what is depicted in the image and (3) non-visual information about the context of the image, such as creator, date, owner, style, etc. The perceptual level is of limited importance for retrieval of paintings and is not easily expressed using ontological concepts. The difference in colour between two paintings, for example, is better described by a histogram than by ontological concepts and relations. Therefore, we will not go into the perceptual level at this moment. In other domains, such as the domain of cellular components in Chapter 3, the perceptual level is more relevant. In that case the Mpeg-7 ontology is a good candidate to cover the perceptual level. The conceptual level is highly important for annotation in the painting domain, but there is little research available that covers this level. VRA, for example, summarises what is depicted in an image in two slots called subject and description. This is not sufficient for the more structured annotations and queries we intend to facilitate, such as: "Woman holding child" or "Man looking at sky". The non-visual level is of equally high importance, and is well covered in literature. Fifteen of the seventeen VRA elements regard this level. Using existing vocabularies is desirable, since it increases the shareability of our annotations. For those reasons, our metadata schema uses the well-known VRA elements for the art-historic descriptions and extends VRA for content descriptions. This is in line with the intentions of VRA, that says

> "The VRA Core 3.0 is intended as a point of departure not a completed application. The elements that comprise the Core are designed to facilitate the sharing of information about works and images among visual-resource collections. These elements may not be sufficient to fully describe a local collection and additional fields can be added for that purpose." (VRA 2002)

For visualisation purposes, the 15 art-historic VRA data elements were grouped into three sets:

Production-related descriptors: title, creator, date, style/period, technique[3], culture and relation.

Physical descriptors: material.medium, material.support, measurements, type and record type.

[3] In Chapter 2 the VRA elements type and technique are considered to be at the perceptual level since they can be derived from the visual characteristics of an image. In this chapter, however, we consider type and technique to be at the non-visual level, since in the domain of paintings they are associated to creator, style and material more then to what is depicted in a painting.

Administrative descriptors: location, collection ID, source and rights.

Two VRA data elements are not included in the metadata schema: description and subject. Both are used to describe the content of the image. As we were interested in providing a more structured content description, we added properties to describe the content of an image in terms of the scenes and objects that are depicted. Scene descriptions comprise events (what is happening in the picture as a whole), places (where is the scene located) and times (when does the scene take place). For descriptions of objects that are depicted in the image, we used an adapted version of the 'sentence structure' proposed by Tam (2002) as a means of structuring the descriptions. The objects that are depicted in an image are described with a collection of statements of the form `agent - action - object - recipient`. Each of these statements should at least have an agent (e.g. a portrait) or an object (e.g. a still life). The concepts used in the sentences are selected from the various ontologies. Multiple sentences may be used to describe a single painting. The tool also provides a free text field, where information can be stored that doesn't fit into one of the slots, or is not present in any of the ontologies. As an example, the painting by Chagall in Figure 5.2, in which Chagall kisses his wive and receives flowers from her, can be described with the following statements (source of the term parenthesised):

Agent: "Chagall, Marc" (ULAN)
Action: "kiss" (WordNet)
Recipient: "wives" (AAT)

Agent: "woman" (WordNet)
Action: "give" (WordNet)
Object: "flower" (WordNet)
Recipient: "Chagall, Marc" (ULAN)

Event: "birthday celebration" (Iconclass)
Place: "artists workplace" (WordNet)

The 'sentence structure' avoids the problems of parsing natural language descriptions, while maintaining some of the naturalness and richness. The naturalness is limited, as can be seen from the plural form wives in the first statement. AAT uses the plural form for concepts whereas Wordnet, Iconclass and ULAN use singular terms.

5.3.3 Linking the metadata schema to the ontologies

Where possible, a slot in the metadata schema is bound to one or more relevant subtrees of the ontologies. For example, the VRA slot style/period is bound to two subtrees in AAT containing the appropriate style and period concepts. The following VRA data elements are currently linked to parts of AAT: technique, style/period, type, record type, material.support, material.medium and

culture. One VRA data element is linked to ULAN, namely creator. The slots of the content description are also linked to subtrees of the ontologies. WordNet provides many general concepts for content descriptions. AAT also provides some useful concepts for this purpose. There is some overlap between AAT and WordNet. In the next subsection we return to this issue. Iconclass is particularly useful for describing scenes as a whole (see the 'birthday celebration' example earlier). ULAN contains specific persons, which are typically used to annotate images in which artists themselves are depicted (e.g., a self portrait). We are currently considering to also include geographical terminology, such as the Thesaurus of Geographical Names (TGN) (The Getty Foundation, 2006c), to be able to describe specific locations in a semantically meaningful way.

5.4 Links Between Ontologies

The five ontologies contain many terms that are in some way related. For example, WordNet contains the concept wife, which is in fact equal to the AAT concept wives (AAT uses the plural form as the preferred one). Without relations between the two concepts this could lead to problems if a painting is annotated with WordNet wife and queried with AAT wives. One could consider designing a new ontology by merging them. However, the general philosophy behind the semantic web is to reuse existing ontologies as much as possible. Therefore, a better option is to use the ontologies 'as-is' and create separate corpora of ontology links. We added three types of ontology links: equivalence links, subclass links and links specific to the painting domain. Equivalence relations and subclass relations are often mentioned in the literature as useful link primitives (e.g. Niles and Pease 2003). They increase recall since they make it possible to retrieve images annotated with concepts from one ontology while searching for concepts from another ontology. Domain-specific links are used to improve recall and speed up annotation.

5.4.1 Equivalence links

We added equivalence relations between terms appearing in multiple ontologies that refer to the same concept. For example, the artistic movements branch in WordNet is linked to the equivalent styles and periods subtree is AAT. Similarly, the WordNet concept wife is linked to the AAT concept wives. As RDF Schema does not provide an equivalence relation,[4] we had to introduce our own special-purpose property for this. In forthcoming versions of the tool we intend to replace this relation with the OWL language construct owl:equivalentClass (Web Ontology Working Group 2003).

5.4.2 Subclass links

When differences between the structures of ontologies are large, which is common, equivalence relations are only possible at the lowest, most specific branches of the hierarchies. We use the

[4]The revised version of RDF Schema (Brickley and Guha 2004) allows cycles of subclass relations. This means that one can now represent equivalence of A and B by stating the A is a subclass of B and that B is a subclass of A.

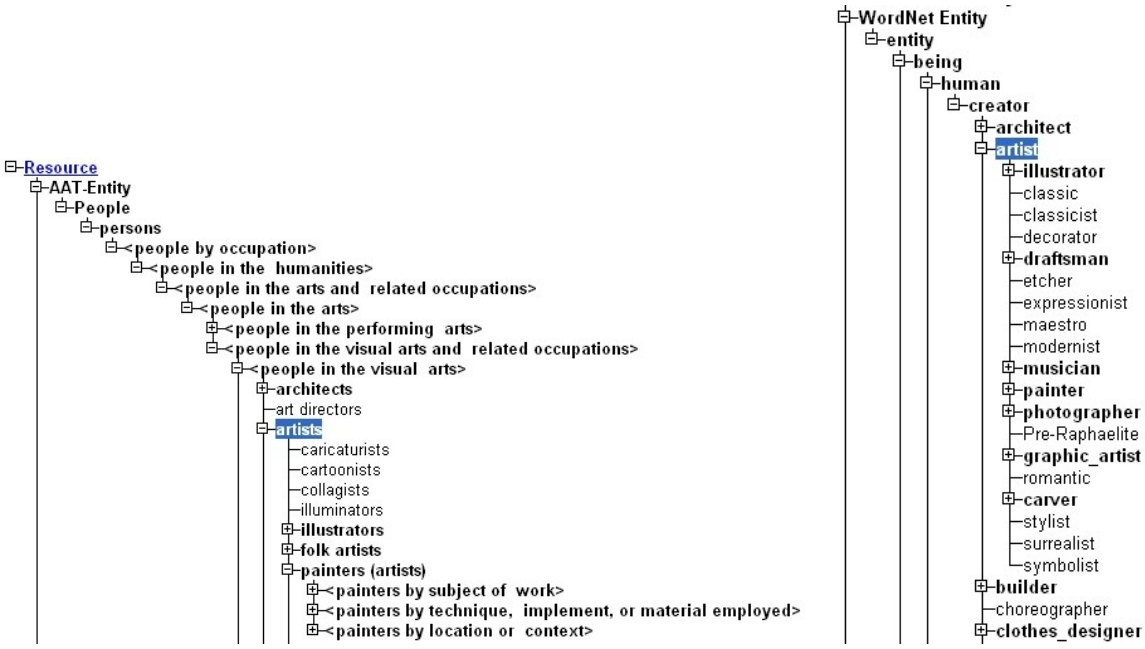

Figure 5.3 Subtrees of AAT (left) and WordNet (right) in which the concept artist appears.

RDFS subClassOf relation to create links at a higher level in the hierarchies. Consider the example in Figure 5.3 which shows a subtree of the AAT and one of WordNet. One can see that the term artist in WordNet does not refer to the same concept as artist in AAT, since some subconcepts of artist in WordNet, such as musician, are not subconcepts of artists in AAT, which contains only people in the visual arts. To link WordNet to AAT we need to create a subclass link: AAT artist is a subclass of WordNet artist.

5.4.3 Domain-specific links

In addition to equivalence and subclass links, we also use relations specific to the domain of paintings. By linking painting techniques to materials, for example, we were able to derive the value of the technique slot from the values of the material.support and material.medium slots. Similarly, a link between artists in ULAN and painting styles in AAT made it possible to offer the user suggestions regarding the value of the style/period slot once the creator was entered. Values in the annotation that are suggested by the tool reduce the time and effort spent by the human annotator. Relations between painters and styles can also improve recall; if a user is searching for Fauvist paintings, the tool can retrieve paintings by Matisse, Derain and De Vlaminck, who are all Fauvist painters. A number of this type of handcrafted links is available in the Mia system: Picasso is linked to cubism, Matisse is linked to Fauve, Van Gogh to impressionism, and so on. This relation is many-to-many: an artist may belong to multiple styles.

Other derivations are possible, but are not yet supported by the tool. ULAN contains information about the country of origin of the artists. This means that the VRA slot culture could in

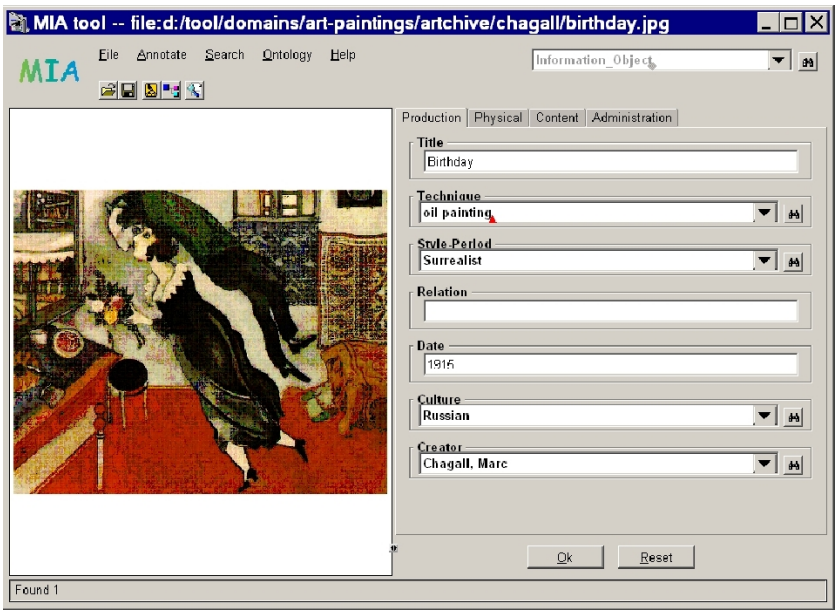

Figure 5.4 Screenshot of the annotation interface, showing the production-related descriptors of the VRA element set.

principle be derived from the slot creator. The type of a painting can sometimes be derived from descriptions of the content. If the only description of a painting is an agent, the painting is probably a portrait. If the agent is equal to the creator, we are looking at a self-portrait. The suggested values act as default values and can be overridden by the annotator.

5.5 Annotation and Search Scenario's

5.5.1 Annotating art-historic features

Figure 5.4 shows a screenshot of the annotation interface. In this scenario the user is annotating an image representing the painting 'Birthday' by Chagall. The figure shows the tab for production-related VRA data elements. The four elements with a binoculars icon are linked to subtrees in the ontologies. For example, if we would click on the binoculars for style/period the window shown in Figure 5.5 would pop up, showing the place in the hierarchy of the concept Surrealist. We see that it is a concept from AAT. The grouping terms of the AAT subtrees from which we can select a value for style/period are shown with an underlined bold font.

5.5.2 Speeding up annotation

In Section 5.1 we mentioned that textual annotations that were present in the Artchive collection were imported in the demonstrator to speed up the annotation process. In addition, in Section 5.4.3 we discussed how the burden on the annotator can be lightened by suggesting values for annota-

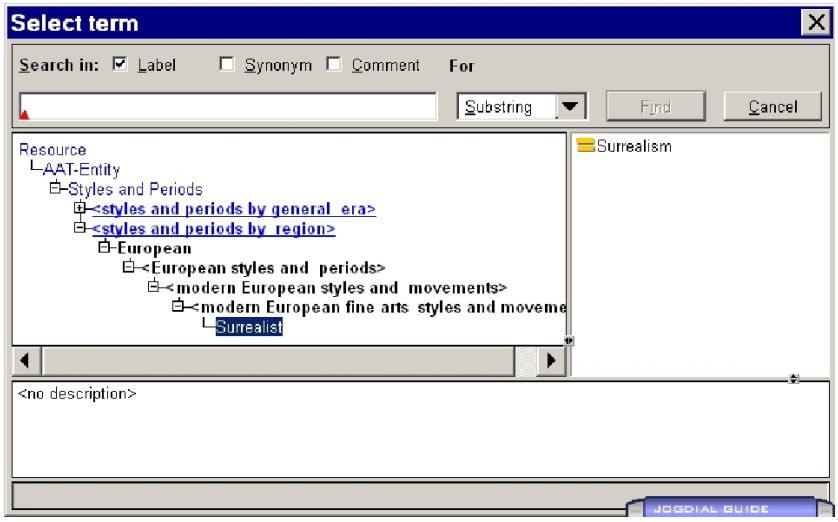

Figure 5.5 Browser window for values of the style/period slot. The concept Surrealist has been selected as a value.

tion, derived from known values. For the style/period slot in Figure 5.4 a value was suggested based on the slot value for creator. The same was done for technique, which can be derived from the two material slots. All values in Figure 5.4 except for culture were derived automatically from the existing annotation.

5.5.3 Annotating content

Figure 5.6 shows the annotation of the content of the painting 'Portrait of Derain' by Maurice de Vlaminck. The schema on the right-hand side implements the content schema as described in Section 5.3.2. The content has been tersely described with the following terms:

Agent: "Derain, Andre" (ULAN)
Action: "smoke" (WordNet)
Object: "pipes(smoking equipment)" (AAT)

As with the art-historic features, the slots are linked to one or more subparts of the underlying ontologies, which provide the concepts for this part of the annotation. For example, if we would click on the binocular icon for `action` the window shown in Figure 5.7 would pop up, showing the place in the hierarchy of the concept smoke. We see that it is a concept from WordNet.

5.5.4 Finding the right term

The ontology browser interface of the Mia tool provides support for finding the right concept. The user can enter a term and the system will provide a popup list of concepts matching the input string. In the popup in Figure 5.7, one synonym of smoke is provided, namely smoking. The

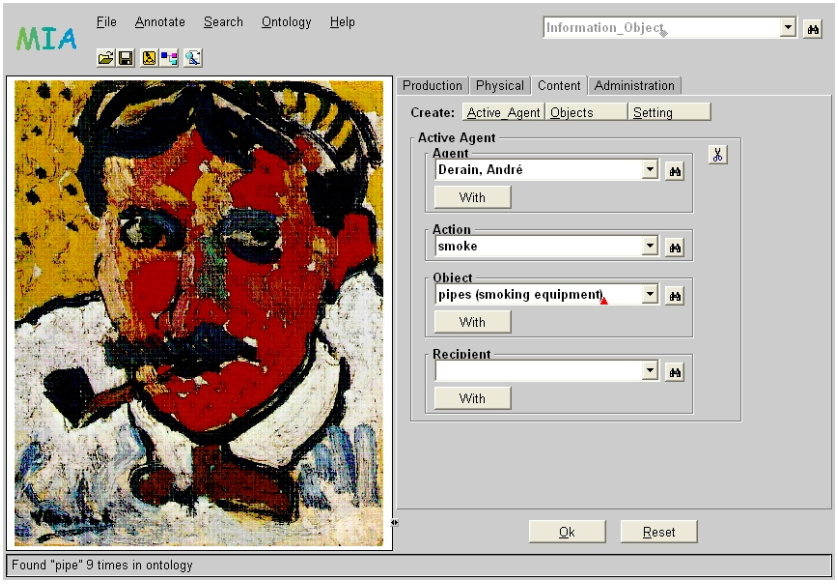

Figure 5.6 Content description of the painting 'Portrait of Derain' by Maurice de Vlaminck, containing an agent, action, object, but no recipient.

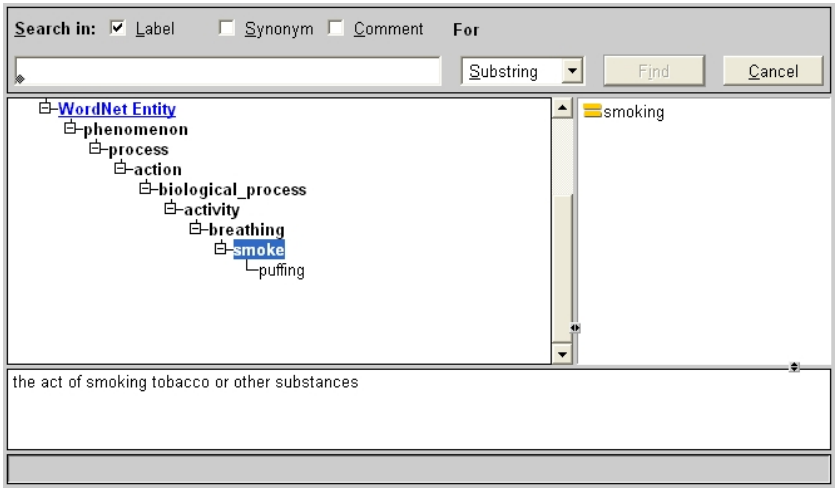

Figure 5.7 Browser window for the concept smoke.

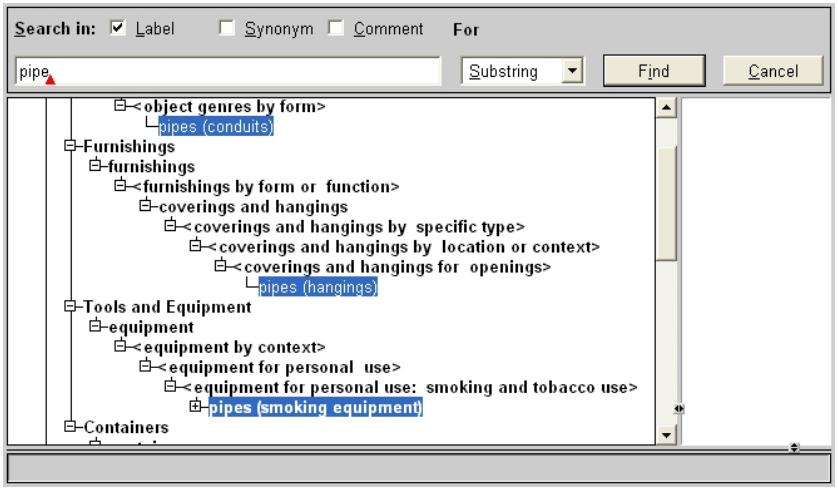

Figure 5.8 Browser window for the concept pipe.

user can now select the concept smoke or, seeing the specialisation puffing, decide to use a more specific concept. In general, we advise annotators to make annotations as specific as possible, as long as the concepts can be traced back to a more general parent.

When a homonymous term is entered, the tool will indicate that this is an ambiguous term. The popup window in Figure 5.8 provides the user with a choice of concepts from the ontologies that are associated with the ambiguous term pipe. It shows three of the concepts associated with pipe, namely conduits, hangings and smoking equipment. From the placement of the terms in the respective hierarchies, it becomes clear to the indexer which meaning of the term is the intended one. The need for term disambiguation occurs frequently when large vocabularies are used.

The ontologies provide a wide range of concepts for content descriptions. Although the choice of concepts depends on the goal and background of the annotator, there are some general guidelines for good annotations. An annotation is most effective if the annotator chooses the concepts as specific as possible. In addition, an annotator should focus on agent and object descriptions when annotating the content of an image, as experiments have shown (Chapter 2; Jörgensen 1996) that users (including searchers) describe images in terms of the agents and objects that are depicted.

5.5.5 Searching for an image

Annotation with concepts from an ontology allows one to perform semantic matching during search. An image can be found by searching for a synonym or hyponym of a concept used in the annotation. For example, the painting ‘Birthday’ in Figure 5.2 was annotated with kiss and can be found by searching for touch. We will return to this issue in Chapter 6.

The tool provides two types of semantic search. With the most simple search option the user can search for concepts occurring in any slot in the image annotation (Figure 5.9). A search for Derain will return paintings by Andre Derain but also the painting in Figure 5.6 by Maurice de

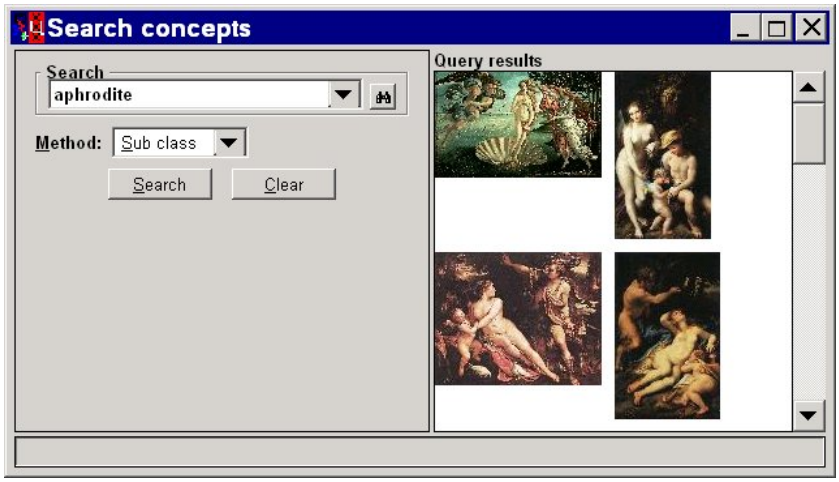

Figure 5.9 Example of concept search. Only a small fragment of the search results is depicted.

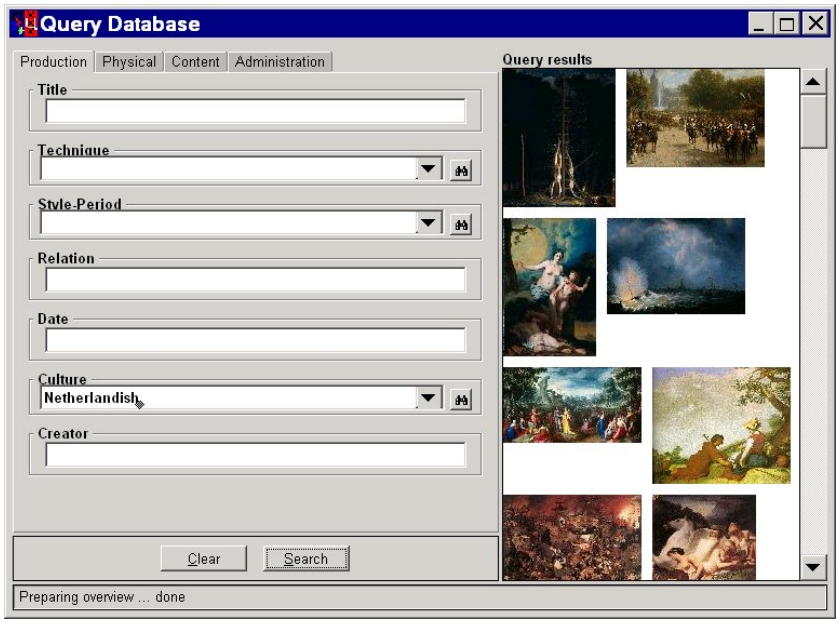

Figure 5.10 Search using the metadata schema.

Vlaminck depicting Derain. The second search option retrieves images which match the query in some part of the annotation. It allows the user to exploit the metadata schema for search proposes. An example of this is shown in Figure 5.10. Here, the user is searching for images in which the slot culture matches Netherlandish. This query retrieves all images with a semantic match for this slot. This includes images of Dutch and Flemish paintings, as these are subconcepts of Netherlandish, but it does not include images that depict The Netherlands.

In Section 5.4 we discussed how links between concepts improve recall. Suppose the user wants to search for images associated with the concept Aphrodite. Because the ontologies contain an equivalence relation between Venus (as a Roman deity, not the planet) and Aphrodite, the search tool is able to retrieve images for which there is no syntactic match. For example, if we would look at the annotation of the first hit in the right-hand part of Figure 5.9, we would find Venus in the title ('Birth of Venus' by Botticelli) and in the content description (Venus (a Roman deity)). The word Venus in the title can only be used for syntactic matches since we do not have an ontology for titles, but the concept in the content description can be used for semantic matches, thus satisfying the Aphrodite query.

5.6 Discussion

This chapter gives some indication on how the semantic web might be applied to image retrieval. Semantic annotation allows us to perform concept search instead of keyword search. It also paves the way for more advanced search strategies. For example, users can specialise or generalise a query with the help of the concept hierarchy when too many or too few hits are found.

In Schreiber et al. (2001) a qualitative analysis was done on the added value of concept search over keyword search. The provisional conclusion was that for some queries (e.g., ape) keyword search does reasonably well, but for other queries (e.g., great ape) the results are suddenly poor. This is exactly where semantic annotation could help.

Although our approach to a large extent relies on manual annotation, we have shown it is possible to generate at least partial semantic annotations from existing annotations (which vary from free text to structured database entries). The application scenario in Section 5.5 shows an example of this. However, the example is based on a special-purpose parser. Systematic use of natural language processing techniques should be considered here. Also, content-based image analysis techniques could be used to derive perceptual-level image annotations, such as the location and colour of objects.

Our experiences with RDF Schema were generally positive. We made heavy use of the metamodeling facilities of RDF Schema, which allow one to treat classes as instances of other classes, for defining and manipulating the metamodels of the different ontologies. In our experience this feature is in particular needed in cases where one has to work with existing representations of large ontologies. This is a typical feature of a semantic web: one has to work with existing ontologies, even if one disagrees with some of the design principles of the ontology. For our purposes RDF Schema has some limitations in expressivity. We especially needed a notion of property cardi-

nality and of equivalence between resources (classes, instances, properties). For this reason we plan to move to OWL, the Web Ontology Language currently under development at W3C (Web Ontology Working Group 2003).

Acknowledgments

We gratefully acknowledge the contributions of Marcel Worring, Giang Nguyen and Maurice de Mare.

6

Query Expansion for Image Content Search

In the previous chapter, we have presented application scenarios in which ontologies aid annotation and search in image collections. In this chapter we continue this line of research by investigating the use of WordNet for image content search. We discuss a metadata schema and annotation facilities of an annotation- and search-demonstrator. In an experimental setting we investigate query expansion strategies using annotations made with the demonstrator. We conclude on an expansion strategy that gives the best balance between recall and precision.

Part of this chapter has been submitted for publication.

6.1 Introduction

The use of ontologies to improve annotation and search of visual resources has been demonstrated in Chapter 5. It was shown that relations between concepts in an ontology can be used to improve recall of queries: a search for flowers will return paintings annotated with sunflowers, a search for cubist paintings will return paintings annotated with Picasso as the creator. In this chapter we examine the use of semantic relations to improve recall of *content* queries, in particular queries for objects depicted in an image.

We discuss annotation properties of a semantic annotation- and search-demonstrator, developed in the E-Culture project (Amin et al. 2006). This demonstrator supports the creation of annotations, storing annotations in RDF and searching for annotation concepts in a web interface, building on the SWI-Prolog SeRQL engine (Wielemaker 2005). It can be seen as a continuation of the Mia demonstrator described in Chapter 5. In this chapter we elaborate on two annotation features of the E-Culture web demonstrator that are different from the Mia demonstrator: (1) the metadata schema for content-annotation and (2) an explorative feature to speed up the annotation process by suggesting values.

In an experimental setting, we study the use of WordNet relations for image content search, using a collection of paintings that was previously annotated with the E-Culture annotation demonstrator. The collection consists of paintings from the Artchive (Harden 2006), a collection that was also used in Chapter 5. We query the annotated collection, using not only our original query concepts, but also closely related concepts. The assumption is that expanding a query with semantically related concepts will result in additional correctly retrieved paintings. A query for Eating, for example, could result in paintings annotated with banquet, since in WordNet wn:banquet (or

feast) is wn:derivationally_related_to wn:feasting, which is a wn:hyponym of wn:eating.

Hierarchical relations such as hyponym, subclass, narrower term (NT) or is-a, are commonly used for query expansion by annotation and search tools. However, WordNet contains many more relations. Seventeen types of relations exist between synonym sets (synsets), such as hyponym, meronym and antonym (Fellbaum 1998). In order to discover which of these relations lead to the best search results, we pose queries with different levels of expansion and examine the results. Intuitively, the more relations we use to expand the query, the higher recall will be. On the other hand, if too many relations are used, precision may become low. The aim of the present work is to identify which relations give the best balance between recall and precision. We will also look into the effect of combinations of relations and the optimal number of nodes between query concept and annotation concept.

Section 6.2 discusses work related to query expansion. In Section 6.3 we present annotation facilities of the E-Culture demonstrator. Section 6.4 contains the design of the experiment and Section 6.5 the experimental results. We conclude in Section 6.6.

6.2 Related Work

The use of WordNet for query expansion has been studied by the information retrieval community. Mostly, this comprised retrieval of textual documents. In order to use semantic relations for retrieval of textual documents, the correct WordNet sense has to be assigned to words in the text. This is called word sense disambiguation (WSD).

Voorhees (1994) demonstrated that the success of query expansion depends on the length of queries and on the selection of the right synsets. She manually and automatically selected query synsets and expanded these with directly related synsets. In her study, she showed that when query synsets were manually selected, recall improved for short queries, but not for longer queries. When query synsets were automatically selected, query expansion did not improve the results at all. Gonzalo et al. (1998) measured the sensitivity of retrieval performance to disambiguation errors. They manually indexed both queries and documents with WordNet synsets, deliberately introducing errors. They found that indexing with synsets improved search substantially if the word sense disambiguation error was less than 10 %. A disambiguation error of more than 30 % produced no improvement over using just the original terms in the queries and documents.

Moldovan and Mihalcea (2000) developed a method for WSD with 87 % accuracy for nouns, which is within the 30 % error margin. They expanded short queries with words that belong to the same WordNet synset. Expansion led to an increase in precision for queries provided by the sixth Text Retrieval Conference (TREC), while there was no increase for queries posed by users of internet search engines. Smeaton and Quigley (1996) used expansion techniques on image captions. They manually disambiguated words from both queries and captions, and added WordNet synonyms to each word. Retrieval based on these expanded queries and documents gave better results than retrieval based on just the original words.

Although most expansion techniques rely on WordNet synonyms, also hyponyms, hypernyms

and words in the glosses have been used. Liu et al. (2004), for example, expanded queries with synonyms, hyponyms and glosses and found that this improved results over non-expanded queries. They do not report on the accuracy of their WSD method.

Few studies compare the effect of different types of relations. Navigli and Velardi (2003) compared retrieval results of original queries to results of synset queries and to results of three types of expanded queries: (1) expansion with hyponyms, (2) expansion with synsets of disambiguated gloss words and (3) with plain words from the glosses. They posed 24 queries provided by TREC 2001 to Google. Expansion with plain words from the glosses gave the best results (22 % increase over original queries), while the other methods only showed an increase of 1 to 3 % over original queries. They do not report on the accuracy of their WSD method and the effect of this on the results. Sim (2004) retrieved URLs, were a URL containing the exact query word is considered most relevant, followed by a URL with a synonym, a hyponym and finally a hypernym. They found that the optimal weight for each query expansion type is 1.0, 0.8, 0.6 and 0.4 for exact words, synonyms, hyponyms and hypernyms respectively. None of these papers, however, report on the effect of *combinations* of WordNet relations on the results of expanded queries.

The consensus seems to be that WordNet relations improve search only if the correct synsets are used in queries and documents. A number of retrieval systems have emerged that make this condition a realistic one; they facilitate annotation and search with WordNet synsets or with concepts from other ontologies.

The Mia demonstrator and the E-Culture web demonstrator are examples of this type of semantic annotation and search applications. Both enable annotation of images with multiple vocabularies, including WordNet. Another well-known example is MuseumFinland. This web-based system integrates collections of several Finish museums by translating the existing annotations to concepts from a number of ontologies (Hyvönen et al. 2004a). The collections can be searched in a multi-faceted browsing interface or with keywords. Alternatively, users are able to search the collection using a multi-facetted thesaurus browser (Hyvönen et al. 2004c). Sinclair et al. (2005) are building a portal from which collections of cultural heritage institutions can be searched and annotated with concepts from ontologies. The CIDOC CRM (Doerr 2003) is used as a common framework to integrate the different metadata schemas used by the institutions. Bloedorn et al. (2005) annotate images with a domain ontology, which is linked to a core ontology (DOLCE) and a visual ontology (Mpeg-7). Other examples of semantic annotation and search tools are the Semantic Markup Tool of Kettler et al. (2005) and the annotation tool for NASA images of Halaschek-Wiener et al. (2005).

Intuitively, most systems use hyponym, subclass or narrower term (NT) relations to expand queries. Although some systems use more than one type of relation – in MuseumFinland meronyms are used as well as hyponyms – none of them report on the added value of different types of relations for search results.

6.3 Annotation with the E-Culture Web Demonstrator

The main objective of the E-Culture web demonstrator is to employ novel semantic web and presentation techniques to provide better indexing and search mechanisms for the knowledge rich domain of cultural heritage (Amin et al. 2006). The demonstrator uses four vocabularies for annotation and search: WordNet 2.0 (Fellbaum 1998), the Art and Architecture Thesaurus (AAT) (The Getty Foundation, 2006a), the Union List of Artist Names (ULAN)[1] (The Getty Foundation, 2006d) and the Thesaurus of Geographical Names (TGN) (The Getty Foundation, 2006c). All were translated from their native format to RDF/OWL. Details about the translation of WordNet to OWL can be found in Assem, van et al. (2004). The TGN contains around 1,102,000 names and other information about places around the world. Each place record (or subject) has properties such as ID, GPS coordinates and place type (e.g. inhabited place, state capital, river). For short descriptions of WordNet, AAT and ULAN we refer to Chapter 5.

6.3.1 Metadata schema for art-historic annotations

The E-Culture demonstrator distinguishes between art-historic annotations (cf. the non-visual level in Chapter 2), such as the creator, style and date of a painting, and content annotations (cf. the conceptual level in Chapter 2), that describe what is depicted in a painting. The metadata schema for art-historic annotations consists of VRA elements. VRA provides *qualifiers* to more precisely identify the meaning of the elements. Based on experiences with the Mia demonstrator and based on expected use, we included the following VRA qualifiers in the annotation interface: measurements.dimensions and measurements.resolution, location.currentSite and location.creationSite, material.medium and material.support.

VRA does not prescribe ranges for its elements. However, for our collection we wanted to restrict the values of some of the properties to relevant parts of the vocabularies. Adding ranges to the original VRA properties would imply changing a standard, which is a contradiction in terms. Therefore, we defined a new class `ec:Work` to be a subclass of vra:Work. `ec:Work` has a restriction on the values of the VRA properties. For detailed information we refer to the RDF/OWL code of the E-Culture metadata schema in Appendix B. This schema represents the (restricted) way we intend to use VRA elements and the extensions we made to VRA. Slots in the art-historic metadata schema and their ranges are depicted in Table 6.1. The prefixes aat, ulan, wn and tgn refer to the namespaces of the four vocabularies used in the system, and properties with the prefix vra are VRA elements. The art-historic annotation interface is form-based, similar to the interface of the Mia demonstrator.

6.3.2 Metadata schema for content annotations

VRA provides only two elements to describe the content of an image, which is insufficient for our purposes. Therefore, we specialised the VRA element vra:subject with subproperties to provide

[1] The E-Culture web demonstrator uses the full version of ULAN, while the Mia demonstrator used a subset.

Table 6.1 Art-historic annotation properties and the parts of the vocabularies that are bound to them in the E-Culture metadata schema.

	Annotation Property	**Range**
Work	vra:creator	ulan:artist
	vra:material.medium	aat:materials
	vra:material.support	aat:materials
	vra:stylePeriod	aat:styles_and_periods
	vra:culture	ulan:nationality
	vra:location.creationSite	tgn:subject
Image	vra:measurements.resolution	rdfs:literal

Table 6.2 Content annotation properties and the parts of the vocabularies that are bound to them in the E-Culture metadata schema.

	Annotation Property	**Range**
Image	ec:subjectType $\multimap$ vra:subject	aat:visual_works_by_subject_type
	ec:event $\multimap$ vra:subject	wn:nouns + rdfs:literal
	ec:place $\multimap$ vra:subject	wn:location + tgn:subject + rdfs:literal
	ec:time $\multimap$ vra:subject	wn:time_period, rdfs:literal
	ec:Objectdescription	ec:ObjectSentence (see Figure 6.1)

more structure in the content descriptions. Table 6.2 depicts the content-annotation properties and the parts of the ontologies that are bound to them. The prefix `ec` denotes the E-Culture custom made properties. The symbol $\multimap$ means rdfs:subPropertyOf and is used to denote that an E-Culture property is a subproperty of a VRA property. A `subject type` property was defined that specifies whether a painting is a still life, landscape, portrait, group portrait, cityscape, or anything else from aat:visual_works_by_subject_type. Three descriptions were provided to describe the scene depicted in the painting: `event` (what is happening in the picture as a whole), `place` (where is the scene located) and `time` (when does it take place). The ranges of these properties include both concepts from ontologies (classes) and literals.

Object descriptions were provided in the form of a sentence. This ‘sentence structure’ is less restricted and thus more generally applicable then the sentence structure used in the Mia demonstrator. Short sentences of the form ‘subject-verb-object’, or ‘subject-relation-object’ can be formed, where the relation can be a spatial relation, a social relation, or any other kind of relation from the hierarchy under wn:relation. With the sentence structure we can distinguish, for example, between paintings depicting a man left of a child and paintings depicting a man holding a child, which would not be possible with the single VRA element vra:subject. Multiple sentences may be used to describe a single painting. Also partial sentences are permitted, such as “man standing” or “child playing”.

Figure 6.1 shows that the sentence is called `objectSentence`, and has a `subject`, `relation`

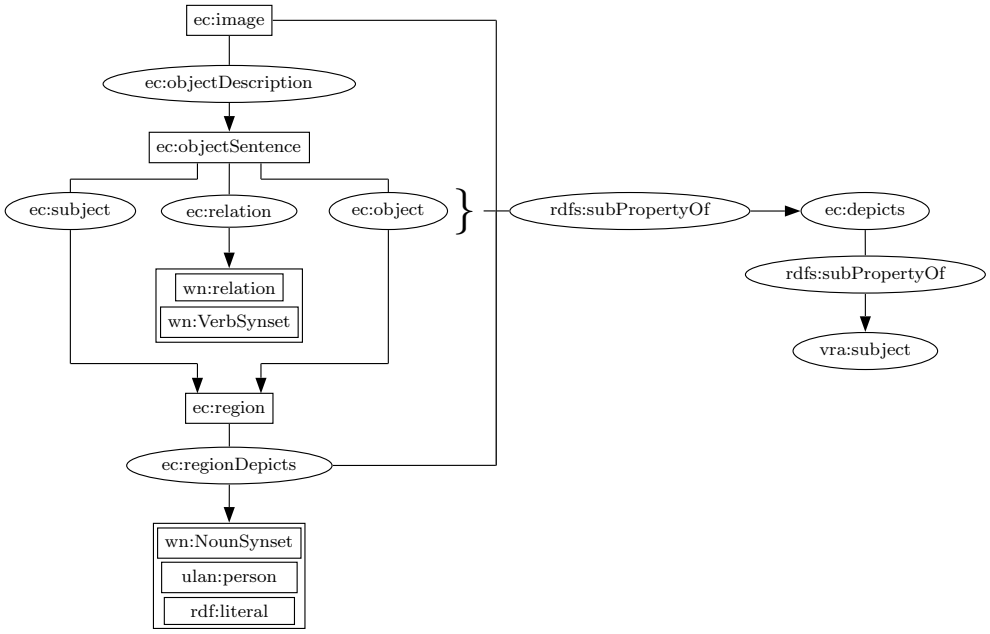

Figure 6.1 Metadata schema for content annotation.

and `object`. The values of `ec:subject` and `ec:object` properties are `ec:regions`, while the value of `ec:relation` is a Wordnet or Ulan concept. A `ec:region` has a `ec:regionDepicts` relation with a WordNet or ULAN concept or with a rdfs:literal. Annotation with literals rather than with concepts from ontologies is a necessary feature for annotation of paintings. Names of specific people (Mme Matisse) or places (Place du Theatre) that are depicted in a painting are often not available in an ontology. Therefore, the metadata schema allows for annotation with concepts from one of the vocabularies as well as with plain text strings for specific concepts. This can also be seen in Table 6.2 in which the range of content properties includes rdfs:literals. Ideally, a specific description in the form of text should only be used in conjunction with a more general concept from an ontology.

In order to ensure that this complex schema can be 'dumbed down' to simple VRA properties, we defined a transitive property `ec:depicts`. The properties `ec:objectDescription, ec:subject, ec:object` and `ec:regionDepicts` are subproperties of this transitive property, which in turn is a subproperty of vra:subject. In this way the relation between an image and its annotation (i.e. a WordNet or ULAN concept) can be 'dumbed down' to a vra:subject relation.

The user interface for object annotations of the E-Culture web demonstrator (Figure 6.2) is graphical, unlike the form-based interface of the Mia demonstrator. The user can create objects and relations between them, and name them with concepts from WordNet, ULAN, or with literals. This graphical rather than form-based approach gives maximal freedom in the number of objects and the number of relations per object. Also, it helps the viewer to identify objects in a painting that are represented in the annotation. It paves the way for future efforts to add additional information

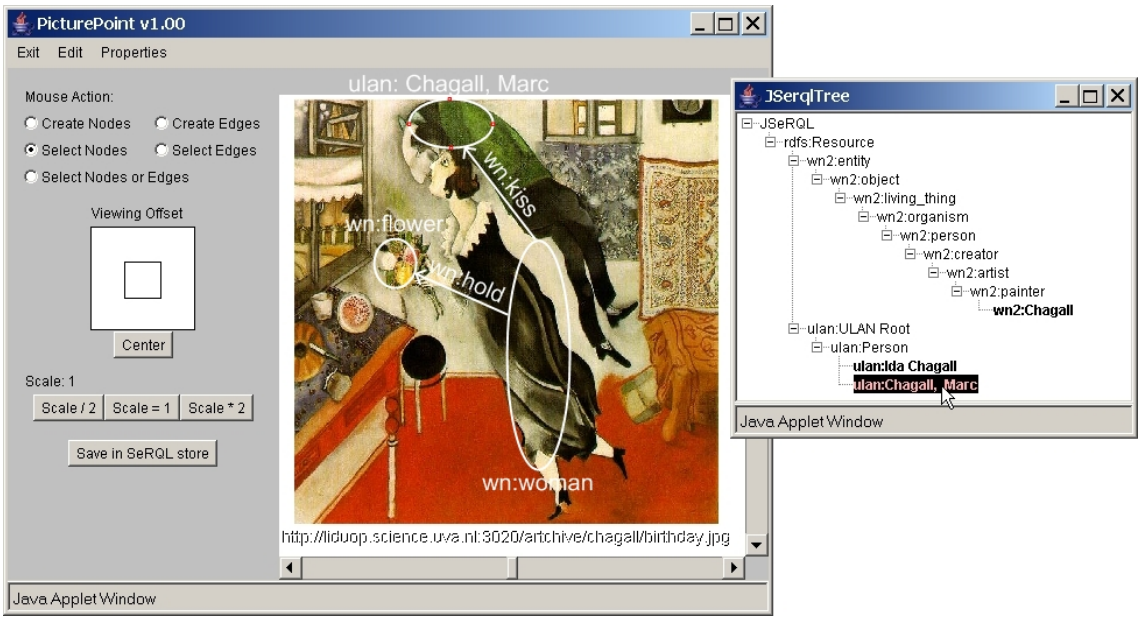

Figure 6.2 Object annotation of Birthday by Chagall.

to the annotation, such as the position and size of objects.

6.3.3 Works and images

The annotation properties in Table 6.1 are divided over `Work` and `Image`. The distinction in VRA between a Work and an Image is highly relevant in the current domain. As pointed out in Chapter 2, VRA defines a work as “a physical or created object”. An image is a representation of the work. It is “the visual surrogate of such [created] objects” (VRA 2002). For paintings, a work is the original work of art on which one can feel the layers of paint, while an image is a photograph of this work, or a digital image of it on the web or in a book. While VRA states that all elements can be used for both Works and Images, we expect that in our domain most VRA elements will be used solely for works. Usually, the creator and date of a painting are more interesting than the creator and date of a photograph of that painting. One VRA element is linked to a (digital) image, namely measurements.resolution. Figure 6.3 depicts the relation between a work an an image in our metadata schema.

Descriptions of the content are also linked to the image instead of to the work. The reason for this is that in almost any annotation system, the annotator is looking at an image, rather than at the actual work. When it is certain that the image depicts the work correctly and completely (and is not, for example, a detail or a black-and-white image of the work), content annotations of images can be interpreted as content annotations of works.

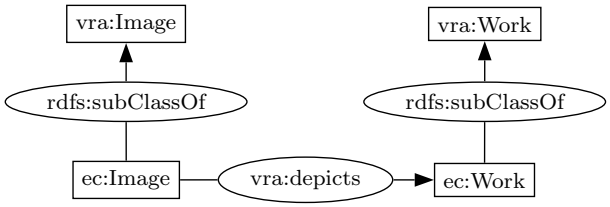

Figure 6.3 Relations between images and works in the E-Culture metadata schema.

6.3.4 Speeding up annotation with suggestions

Creating a detailed annotation is a time consuming process. In Chapter 5, we have shown how this process could be enhanced by deriving values of slots in the annotation if one slot is filled in. An add-on to the E-Culture system makes it possible to derive more values from existing background knowledge. From the value of the creator slot, we derive style, culture and material. Culture is derived from knowledge available in ULAN about the nationality of the artist. Style is derived from links between styles and creators as published by De Boer et al. (2006). In our vocabularies, there is no knowledge available that links artists to materials. However, in our experience, most artists use only two or three materials throughout their career. Therefore, the value of the material slots is derived by looking at the most frequent values in previous annotations of the same creator. This strategy is also used to provide additional suggestions for style and culture. Figure 6.4 shows suggestions for a painting by Chagall. Three values are suggested for `culture`, since according to ULAN Chagall had three nationalities during his life. A user can select correct values and add them to the annotation by clicking 'Apply'.

Currently, the suggestion add-on gives a list of at most 10 suggestions, of which the first 5 are based on previous annotations and the last 5 on links between concepts in the vocabularies. The length and order of the suggestion list is a feature that still needs testing and evaluation. Also, the question of how decisive the suggestions should be is an open issue: should the derived values be filled in automatically, or does the annotator need to confirm them? Other issues that need further research include the situation that multiple derived values are equally likely to be the correct value, and the situation that the annotator fills in values that conflict with derived values.

Despite the many open issues on the matter, the idea of lightening the burden on the annotator by deriving annotations from known annotations has been recognised by the multimedia community. In Ahern et al. (2005) a list of people interested in a certain photo is derived from knowledge about who is present at the moment the photo was taken, the social group of the recipient and the time schedule of the user. O'Hare et al. (2005) derive the time of day, weather and indoor/oudoor data from the location, time and camera settings of the camera when the photo was taken. Lahti et al. (2005) derive annotations of the event, setting and people present in a photo, based on GPS data, time, the calender and the address book of the user.

In principle, we could derive more slots than the ones described above, further limiting the time spent on annotation. From creator and date we could derive location.creationSite if the places

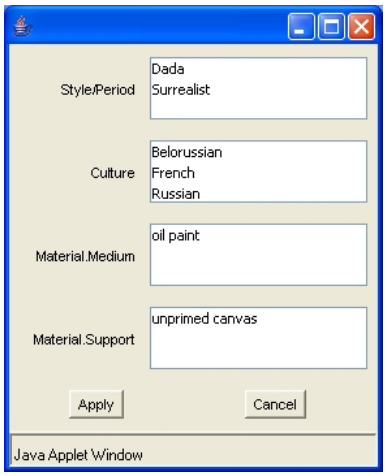

Figure 6.4 Suggested annotation values for a painting with creator Chagall.

of residence of the creator were known. If the only description of a painting is a wn:person or ulan:artist, the painting is probably a portrait. If the subject is equal to the creator, we are looking at a self-portrait.

6.4 Experimental Setup

In order to find out which (patterns of) relations lead to improvements in search results, we queried a collection of Artchive paintings annotated with WordNet synsets. A total of 202 paintings by 25 painters were annotated by 12 members of the E-Culture project (E-Culture 2006). The annotators were given a set of guidelines to ensure a uniform view on content annotation[2]. The annotators were moderately familiar with the vocabulary (WordNet) and were not aware of the research questions to be answered in the present experiment. The resulting annotations and the RDF version of WordNet were stored in a Sesame repository and queried with SeRQL (Broekstra and Kampman 2003).

Fifteen query concepts were chosen by looking at objects depicted in paintings in the Artchive collection that were not annotated or used in the experiment. The query concepts were chosen to be on Rosch's basic level (Rosch 1976) (see also Chapter 2). One exception was the query concept Tree, which is more general then the basic level. In flora and fauna, the basic level is usually on the level of 'genus', which for trees would have been oak or chestnut. The annotators, however, were not able to distinguish a chestnut from an oak, especially in paintings. This justifies the use of the more general query concept Tree. The 15 query concepts are listed in Table 6.3. None of the queries were directly related to each other, although some were related through one or more intermediate nodes. Window and House are both related to wn:building; Hand, Male_child and

[2] http://www.cs.vu.nl/∼laurah/ECultureGuidelines.pdf

Woman are all related to wn:person. Each query was posed in 3 ways:

exact-queries: only paintings that are annotated with the query concept are returned

hyponym-queries: paintings that are annotated with the query concept and paintings annotated with a concept that is related to the query concept through hyponym relations are returned. Up to four intermediate nodes are allowed.

all-relations-queries paintings that are annotated with the query concept and paintings that are annotated with a concept that is in any way related to the query concept are returned. Up to four intermediate nodes are allowed.

Recall and precision of each query was measured by comparing the results to a golden standard of matching paintings for that query concept. To come to a golden standard, all paintings were judged by two raters. Cohen's Kappa (κ) was used to measure correspondence between raters. The mean κ of all query concepts was 0.68, which is acceptable (Carletta 1996).

6.5 Results

Table 6.3 shows the number of relevant paintings in the collection (Rel), the number of retrieved paintings (retrieved), the number of correctly retrieved paintings (correct hits), recall and precision of each query in each condition: exact-queries (Ext), hyponym-queries (Hyp) and all-relations-queries (All). Recall appears to be low for all query types. This is due to the fact that the raters were advised to make the golden standard strict; when a query concept was visible in an image, no matter how small or insignificant, the image was counted as a hit. The annotators, on the other hand, only annotated objects that were clearly visible or important in the image. This frequently led to situations in which raters considered a painting relevant because it depicted an object matching a query concept, but annotators did not annotate the object because it was not important. A painting depicting, for example, an apple and a a bottle, could be annotated with just apple, but counted as a correct hit for both apple and bottle. This had a negative effect on recall. Similarly, it might have had a positive effect on precision. Therefore, the recall and precision values of each query type can only be understood in relation to the recall and precision of the other query types.

One of the fifteen query concepts, Trunk, was left out of the analysis. It received no hits on exact-queries or hyponym-queries and incorporating it would corrupt paired t-test and ANOVA results. It was therefore also not used to determine the mean values in Table 6.3. Nonetheless, Trunk provides a good illustration of the added value of other types of relations than just hyponyms. The fact that Trunk is a meronym of tree made it possible to return all paintings annotated with tree for the query concept Trunk, which lead to high recall (0.37) and precision (0.56).

The three conditions were compared amongst each other with one-way repeated measures ANOVAs. There was a significant effect of query type on recall (F(2, 26) = 46.99, $p < 0.01$). Also, there was a significant effect of query type on precision (F(2, 26) = 63.8, $p < 0.01$). Paired t-tests showed no significant difference between precision of exact-queries and hyponym-queries. There

Table 6.3 Precision and recall of queries over query types.

		Retrieved			Correct Hits			Precision			Recall		
Query	**Rel**	Ext	Hyp	All	Ext	Hyp	All	Ext	Hyp	All	Ext	Hyp	All
mountain	15	6	6	30	5	5	10	0.83	0.83	0.33	0.33	0.33	0.66
window	49	2	2	64	2	2	28	1.00	1.00	0.44	0.04	0.04	0.57
cloud	53	2	4	40	1	3	21	0.50	0.75	0.53	0.02	0.06	0.40
hand	56	3	3	62	3	3	31	1.00	1.00	0.50	0.05	0.05	0.55
male child	4	1	1	66	1	1	2	1.00	1.00	0.03	0.25	0.25	0.50
guitar	4	2	2	19	1	1	3	0.50	0.50	0.16	0.25	0.25	0.75
horse	7	1	1	43	1	1	2	1.00	1.00	0.05	0.14	0.14	0.29
chair	12	5	5	35	5	5	7	1.00	1.00	0.20	0.42	0.42	0.58
woman	59	15	20	52	13	17	38	0.87	0.85	0.58	0.22	0.29	0.64
apple	6	2	2	28	2	2	5	1.00	1.00	0.18	0.33	0.33	0.83
bottle	4	2	2	50	1	1	1	0.50	0.50	0.02	0.25	0.25	0.25
house	37	7	10	53	7	10	22	1.00	1.00	0.42	0.19	0.27	0.59
river	15	6	8	93	5	7	11	0.83	0.88	0.12	0.33	0.47	0.73
tree	49	13	17	30	12	15	17	0.92	0.88	0.57	0.24	0.31	0.35
(trunk	*49*	*0*	*0*	*32*	*0*	*0*	*18*	.	.	*0.56*	*0.00*	*0.00*	*0.37)*
mean	26.43	4.79	5.93	48.50	4.21	5.21	14.41	0.85	0.87	0.29	0.22	0.25	0.55

was a significant difference between precision of exact-queries and all-relations-queries ($p < 0.01$) and between hyponym-queries and all-relations-queries ($p < 0.01$). Paired t-tests showed that recall differed between all query types: between exact-queries and all-relations-queries ($p < 0.01$), between hyponym-queries and all-relations-queries ($p < 0.01$) and between exact-queries and hyponym-queries ($p = 0.017$).[3]

The results showed that expansion with hyponyms of the query concept increases recall, while maintaining the high precision of exact-queries. The use of other types of relations further increases recall but lowers precision, as was expected. The mean increase in recall of all-relations-queries over hyponym-queries was 0.30 ($0.55 - 0.25$). This increase could in part be attributed to the higher number of retrieved images. However, the increase in recall was more than we would expect from the additionally retrieved images only. Suppose that the additional number of retrieved images were randomly taken from the collection, then we would expect an increase in recall of 0.16 according to the following equation[4]:

$$R_{incr.} = \frac{1}{15} \sum_{i=1}^{15} \frac{(Ret\_All_i - Ret\_Hyp_i) \cdot (Rel_i - Hit\_Hyp_i)}{202 - Ret\_Hyp_i} \cdot \frac{1}{Rel_i}$$

where $R_{incr.}$ is the mean expected increase in recall, Ret_All_i is the number of retrieved images by an all-relations-query for query i, Ret_Hyp_i the number of retrieved images by a hypernym-query for query i, Rel_i is the number of relevant images in the collection for query i, Hit_Hyp_i is the number of hits of a hyponym-query for query i and 202 is the total number of paintings in the

[3] In the case of three t-tests with df = 13 and $\alpha = 0.05$, Bonferoni adjustment calls for a significance level p of at most 0.017. None of our p values exceeded this level.

[4] This equation is derived from the equation $\mathbb{E}x = \frac{n \cdot m}{r}$. This problem is also known as the "urn problem", since it asks for the expected number of white balls ($\mathbb{E}x$) out of n balls that are drawn from an urn, containing m white balls and $r - m$ red balls

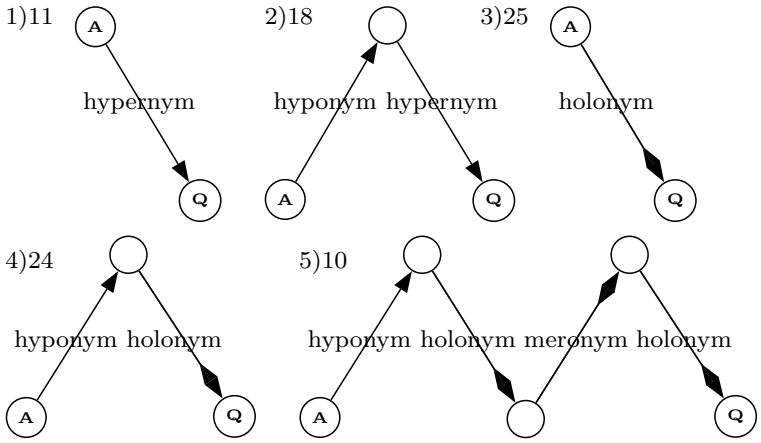

Figure 6.5 Patterns of relations that contributed most to recall and the number of correct hits they produced.

collection. Comparing the increase in recall in our experiment to the expected increase in recall based on additionally retrieved images only, we found the experimental values to be significantly higher ($p < 0.01$).

Examining the results of the hyponym- and all-relations-queries, we found that patterns containing four intermediate nodes between query and annotation (which was the maximum in our experiment) were not beneficial to the results: those patterns led to 231 incorrectly retrieved images and only 25 hits. For example, Monet's 'The Thames below Westminster' was incorrectly returned for the query concept Mountain, since it was annotated with Thames, which is a meronym of - England - holonym of - Pennines - hyponym of - hills - hyponym of - natural_elevation - hypernym of - mountain.

All-relations-queries correctly retrieved 143 paintings that were not found with hyponym queries. The additional hits were caused by 21 distinct patterns of relations (excluding patterns with more than four intermediate nodes). Transitivity of hypernym, hyponym, meronym and holonym relations was assumed to come to the 23 patterns, so hypernymOf - hyponymOf and hypernymOf - hypernymOf - hyponymOf were counted as the same pattern. We interpreted the WordNet relations memberHolonym, substanceHolonym and partHolonym as one type: Holonym. The same was done for different types of Meronym relations. Over 90 % of the Meronyms and Holonyms were partMeronyms and partHolonyms.

The five patterns that led to the most additional hits are depicted in Figure 6.5. Pattern 5 is caused solely by the query concept Hand since WordNet contains the following facts: person - holonym of - body - meronym of - human - holomym of - hand. This caused all paintings of people to be returned for the query Hand. As this structure is present for all body parts, we do not consider this an outlier. Pattern 4 combines two types of relations: hyponym and holonym. An example of a painting that was retrieved by this pattern is 'Wheat Field' by Van Gogh. It contains a house which is a hyponym of - building - holonym of our query concept Window.

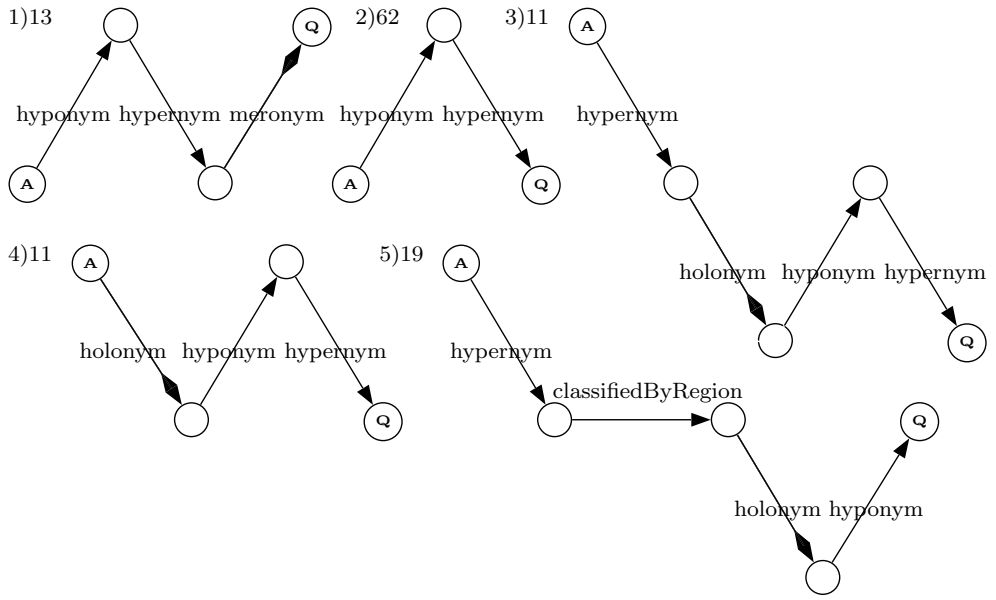

Figure 6.6 Patterns of relations that caused low precision and the number of incorrectly retrieved paintings.

All five successful patterns involve solely hypernym, hyponym, holonym and meronym relations. Other types of relations occurring in various patterns led to few hits while resulting in a considerable number of incorrectly retrieved images. Examples are patterns involving antonym (5 incorrect, no hits), inSynset (7 incorrect, no hits), classifiedByRegion (60 incorrect, 1 hit) and classifiedByTopic (17 incorrect, 3 hits)[5]. Note that the relations InSynset and containsWordSense are not between two synsets, but between words-and-synsets or words-and-word-senses respectively. Relations involving words or word-senses occurred because we did not require intermediate nodes to be synsets. However, these relations were rare and did not lead to any hits. ClassifiedByTopic was useful only for the query concept River, since it links wn:river to wn:body_of_water.

Figure 6.6 shows the five patterns that lead to the highest number of incorrectly retrieved images. Pattern 5, for example, incorrectly returned 'The Empire of Lights' by Magritte for the query concept River, because the painting contains a house and WordNet has the following statements: house - hypernym of - maisonette - classified by region - France - holonym of - Loire - hyponym of - river. Comparison of Figures 6.5 and 6.6 shows that the pattern hyponym-hypernym, also called 'siblings', returns many hits, but even more incorrect images. Siblings are therefore not advantageous for retrieval. Not only siblings, but all other combinations of hypernym with another property (e.g. meronym or holonym) appear disadvantageous. Patterns that combined hypernym with another property led to 154 incorrect images and only 28 hits. Hypernym alone did give good results. Pattern 1 in Figure 6.5 summarises hypernym relations with zero or one intermediate node.

[5]We use the WordNet property names as published in Assem, van et al. (2006). Explanation of the WordNet terminology can also be found in the WordNet manual on http://wordnet.princeton.edu/man/wngloss.7WN

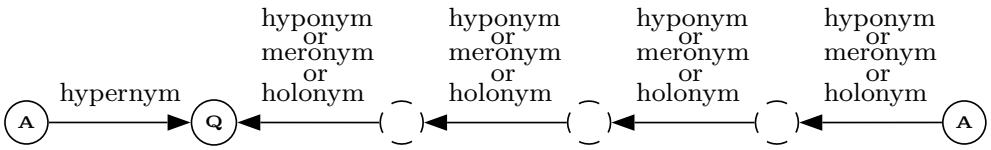

Figure 6.7 Proposed query for expansion. Q is the query concept and A represents the annotation concept. Optional intermediate nodes are dashed.

Longer chains of hypernym relations did not occur in our experiment.

Concluding, it appears that for optimal retrieval results the relation between query concept and annotation concept should be a hypernym relation with up to one intermediate node, or any combination of hyponym, meronym and holonym with up to three intermediate nodes. We propose the pattern in Figure 6.7 to expand queries with.

We expanded the 15 query concepts with the proposed pattern. Table 6.4 shows that the proposed query results in a recall of 42 % and precision of 64 %. The performed t-tests showed a significant difference between hyponym queries (Table 6.3) and the proposed query for precision ($p < 0.01$) and recall ($p < 0.01$). This shows that query expansion with the right types of relations can improve recall with almost 70 % over expansion with only hyponym relations (from 0.25 to 0.42), while preserving an acceptable level of precision.

A good indication of the overall performance of a retrieval strategy is the F-measure. The mean F_1-scores of exact-queries, hyponym-queries, all-relations queries and the proposed queries were 0.33, 0.36, 0.33 and 0.46 respectively. Although a significant increase in F_1 of the proposed query over hyponym-queries could not be proven (t = -1.95, df = 13, p = 0.07), the numbers indicate that the proposed query performs better than the other expansion strategies.

6.6 Discussion and Conclusion

In this chapter we first discussed annotation features of a semantic annotation- and search-tool, and subsequently used semantic annotations created with this tool to study the effect of query expansion on image content search.

We proposed a metadata schema for annotation of both art-historic features and image content. Image content annotation is different from art-historic annotation and requires a different kind of metadata schema. The art-historic features of an image are well understood. A form-based interface, in which users enter annotations in pre-defined metadata slots, is therefore suitable for art-historic annotations. Content annotations, on the other hand, are less well-defined. In Chapter 2 we showed that one painting can be described by numerous classes at several levels of abstraction. Representing all these possible annotation types in a form-based manner would lead to a complex user interface. As an alternative to the content-annotation schema in the Mia demonstrator (Chapter 5), we developed a less structured content-annotation schema that allows a possibly large number of descriptions of the form `subject - relation - object`, without prescribing the

Table 6.4 Precision and recall of the proposed query.

Query	Rel.	Retrieved	Hits	Precision	Recall
mountain	15	8	6	0.75	0.40
window	49	40	24	0.60	0.49
cloud	53	22	15	0.68	0.28
hand	56	38	19	0.50	0.34
male_child	4	5	1	0.20	0.25
guitar	4	9	3	0.33	0.75
horse	7	1	1	1.00	0.14
chair	12	12	6	0.50	0.50
woman	59	28	23	0.82	0.39
apple	6	5	5	1.00	0.83
bottle	4	7	1	0.14	0.25
house	37	18	16	0.89	0.43
river	15	9	7	0.78	0.47
tree	49	20	16	0.80	0.33
(trunk	*49*	*19*	*17*	*0.89*	*0.35)*
mean	26.43	15.86	10.21	0.64	0.42

type of descriptions.

We have shown two ways of using background knowledge: generating annotation suggestions and query expansion. We suggested a prototype 'suggestion tool' that extends annotations of the creator of a painting with other art-historic annotations, thus shortening the time required to produce a complete annotation. The suggestions were based not only on links in ontologies between the creator and other properties, but also on existing annotations. This means that the more paintings are annotated, the better the suggestions become. Content annotations are less suitable for this type of '*annotation expansion*'. Because of the many interpretations of one image, a content annotation can be expanded in so many directions, that it becomes infeasible to suggest to an annotator all potentially meaningful expansions. More promising for content search is '*query expansion*'. We examined the use of various WordNet relations and concluded on patterns of relations that proved most beneficial for query expansion.

Expanding queries with hyponyms is intuitive and frequently used by search tools. The present study showed that it indeed improves recall while maintaining precision. The results also show that recall of retrieval results can be further improved if other types of relations are used as well. Expansion with a combination of hyponym, holonym and meronym relations improves recall while maintaining an acceptable level of precision. Likewise, expansion with hypernym relations improves search results. However, a combination of hypernyms with other types of relations (e.g. hyponyms or holonyms) is more detrimental to precision than it is beneficial to recall. Expansion with other types of WordNet relations, such as inSynset and classifiedByRegion, appeared to harm the results. We can conclude that semantic annotation and search systems such as the Mia demonstrator and the E-Culture web demonstrator can improve their recall values by expanding

query results with not only hyponym relations, but also with holonym, meronym and hypernym relations.

The results of the present study can also be used to improve ranking of result sets. Images linked to a query concept through a successful pattern would then appear higher in the result list than images linked through a less successful pattern. Images linked through, for example, the 'all-relations-query' would end up at the bottom of the list. For queries that yield very little results, expansion with patterns that cause low precision might still be advantageous.

In the present experiment, the gain in recall caused by expansion was considerably higher than what was found by the text retrieval community (see Section 6.2). This might be due to the fact that in our application both queries and annotation were short, often consisting of only a few concepts. Voorhees (1994) found that the effect of expansion is higher for short queries than for long queries. The same might hold for the length of annotations. This suggests that query expansion is especially fruitful for image retrieval, that typically involves short documents (annotations) and short queries.

The results of the experiment are not specific to the E-Culture demonstrator, since the experiment did not rely on this system, other than for collecting the annotations. Also, we expect that the results can be generalised to domains other than the painting domain. The annotations consisted mostly of everyday concepts that occur in numerous other domains, such as news, collections of photographs, movies, etc. However, we cannot assume that the results can be generalised to vocabularies other than WordNet. The specific structure of WordNet, such as the depth of the hierarchy and the frequency of certain types of relations, cannot be separated from the patterns of successful relations that were the outcome of this study.

The depth of the hierarchy varies greatly in WordNet. Most parts of the hierarchy are relatively shallow, but some parts, such as the hyponym hierarchies of flora and fauna, are more than 14 levels deep. In our experiment, we limited the depth of relations between query concept and annotation concept to four intermediate nodes. We found that patterns containing four intermediate nodes performed worse than those containing up to three intermediate nodes. However, we suspect that in some cases a deeper approach of up to seven or eight intermediate nodes will make a positive difference on retrieval of concepts in deep hierarchies such as plants and animals. In our study, for example, Apple could not be related to Plant because they are related with more then four intermediate nodes. An alternative strategy that needs additional testing is to allow hyponym relations to have an arbitrary depth. Some databases interpret hyponym relations as a transitive property and pre-compute the complete transitive closure. In those cases, the length of the pattern of hyponym relations will not cause the query to be computationally expensive. The Mia demonstrator applies this strategy.

Acknowledgements

We would like to thank Alia Amin, Mark van Assem, Michiel Hildebrand, Zhicheng Huang, Janneke van Kersen, Borys Omelayenko, Jacco van Ossenbruggen, Novan Pavlovich, Guus Schreiber,

Ronny Siebes, Jan Wielemaker and Bob Wielinga from the E-Culture team for creating the annotations. Special thanks go to Jan Wielemaker, Ronny Siebes, Jacco van Ossenbruggen, Victor de Boer, Mark van Assem and Alia Amin for their valuable contributions to the demonstrators, the vocabularies and the experiment. Thanks to Alistiar Vardy for rating the images to produce the golden standard.

7

Adding Spatial Semantics to Image Annotations

Chapters 5 and 6 discussed how ontologies can be used to support the creation and use of semantic annotations. In this chapter we explore how content-based image retrieval can be used to automate part of the annotation process. Existing semantic annotations of objects depicted in a painting are extended with spatial information regarding the absolute and relative positions of the objects. This will reduce the time spend by the annotator, as in Chapter 2 we found that spatial information is frequently used in descriptions of images.

This chapter is a slightly modified version of a paper that was presented at the International Workshop on Knowledge Markup and Semantic Annotation at ISWC (Hollink et al. 2004a), and was co-authored by Giang Nguyen, Guus Schreiber, Jan Wielemaker, Bob Wielinga and Marcel Worring.

7.1 Introduction

Making a complete and elaborate annotation of the content of an image is a time consuming process. Therefore, the human annotator should be supported in this task as much as possible. Despite improvements in the field of content-based image retrieval (CBIR), fully automatic annotation of images has not yet reached a satisfactory level of precision. This is mainly due to the fact that the interpretation of what is depicted in an image is subjective. Semi-automatic annotation, however, in which user-generated annotations are complemented with automatically-generated annotations, has the potential to reduce the time spent by the annotator, while maintaining an acceptable level of precision. In this chapter we explore this type of semi-automatic annotation: semantic annotations are augmented with automatically derived spatial information. An image containing, for example, two annotated objects ‘table’ and ‘lamp’ could be augmented with the information that ‘the table is on the right side of the image’ and ‘the lamp is above the table’.

Spatial information is relatively objective, which makes it a good starting point for an exploration into semi-automatic annotation. In addition, in an earlier study (Chapter 2) it was shown that people who describe images often use spatial descriptions. Spatial information is important for describing the composition of an image.

We use a collection of paintings from the Artchive (Harden 2006) that was also used in Chapter 5 and 6. Objects that are visible in the paintings were annotated with concepts from WordNet or AAT using the Mia Demonstrator discussed in Chapter 5. We developed a semi-automatic annotation-tool that extends these semantic annotations with spatial information. The system uses

a vocabulary consisting of concepts from WordNet and the Suggested Upper Merged Ontology (SUMO) to represent spatial properties. We present the results of a small evaluation study in which annotations generated by the system are compared to manual annotations by humans.

This work can be seen as an exploration into bridging the 'semantic gap' (Smeulders et al. 2000), which refers to the cognitive distance between the analysis results delivered by state-of-the-art image-analysis tools and the concepts humans look for in images. It is an exploratory study to investigate the potential of a collaboration between semantic annotation and content-based techniques for image annotation at a conceptual level. The system should be seen as a prototype and the spatial vocabulary as a first step towards a representation of spatial information.

In section 7.2 we discuss work related to the representation of objects and spatial properties, the use of spatial information in search and the qualitative definition of spatial properties. Note that this discussion is exhaustive on none of these topics. Section 7.3 discusses the representation of spatial information. In Section 7.4 we describe the semi-automatic system. Section 7.5 contains the setup and results of a small evaluation study. Section 7.6 contains a discussion.

7.2 Related Work

Talmy (1983) describes spatial relations in the context of human perception. He conveys that the spatial disposition of an object in a scene is always characterised in terms of another object. The first object, which is called the 'figure', is the subject in the expression. The second object, or the 'ground', is used as a fixed reference to which the position of the figure is described. Grounds are for example the earth or the body of the speaker. More then one ground object is possible. In the expression "the bike is on the other side of the church", for example, the bike is the figure, the church is the ground object and the body of the speaker is the second ground object. Another important point is that in human language a finite number of words is used to represent an infinite number of spatial configurations. This means that choices have to be made about which spatial concepts are used in a vocabulary. Cohn (2001) points out that when making a representation of space, questions have to be addressed regarding the kind of spatial entity being used (e.g. regions, points) and the way of describing properties and relationships of these entities (e.g. their topology, size, distance, orientation or shape). We will elaborate on the choices we made regarding the spatial concepts in our vocabulary and the representation of objects in section 7.3.

Lee and Hwang (2002) developed a tool in which users can select regions in images and tag them with keywords. The selected regions are represented as points, and the spatial positions and relations of the regions are computed by looking at the x and y coordinates of the points. We propose a similar division of work: objects are annotated by a user, spatial information is generated by the tool. However, we intend to embed the spatial information more thoroughly into the semantic annotation by using spatial concepts from existing ontologies. Moreover, we want to provide more intuitive representations of spatial concepts than the representation used by Lee and Hwang. Therefore, we use the theory of Abella and Kender (1999), who presented a framework for the representation of spatial information that matches a user's expectations. They provided

NW	N	NE
W	C	E
SW	S	SE

Figure 7.1 Absolute Spatial relations.

quantitative definitions for qualitative spatial properties as expressed by users, such as far, near, above and below. They proposed a theory to derive these properties from images, using mainly the bounding box of objects and the centre point of the bounding box.

7.3 Representation of Spatial Concepts

For our practical purpose of annotating objects in images, we restrict ourselves to two-dimensional, binary relations between regions. The spatial concepts that are included in our vocabulary must be (1) relevant for image annotations and (2) suitable for automatic detection. This last requirement disqualifies concepts like 'behind' and 'in front of' since they are often hard to detect, especially in the painting domain. In order to use the spatial information as semantic annotations, we used concepts from existing ontologies.

There are two types of spatial concepts: absolute positions and spatial relations. The first are used to describe the position of objects within an image. The image functions here as the 'ground'. A common representation of absolute positions are the compass points north, south, east, west, north-east, south-east, north-west and south-west. We divided an image into nine squares where each of the outer squares represents one of the compass points and the middle square represents the centre (Figure 7.1). All absolute positions were taken from the general lexical database WordNet (Fellbaum 1998).

Spatial relations are used to describe positions of objects relative to each other; one object is the 'figure', the other is the 'ground'. The relations that we used in this chapter are: right, left, above, below, near, far and contains. One additional spatial relation can be derived, namely next is either left or right. Spatial relations were taken from SUMO (Niles and Pease 2001). This is a large, well structured ontology that takes into account Cohn's ideas about spatial relations.[1] One exception was the spatial relation far that was taken from WordNet since it was not a concept in SUMO (version 1.15).

For each spatial relation we specified whether it is a symmetric relation, whether it is a transitive relation and what the inverse Of the relation is. RDF/OWL was used for the representation of

[1]CVS log for SUO/Merge.txt, http://ontology.teknowledge.com/cgi-bin/cvsweb.cgi/SUO/Merge.txt, revision 1.24

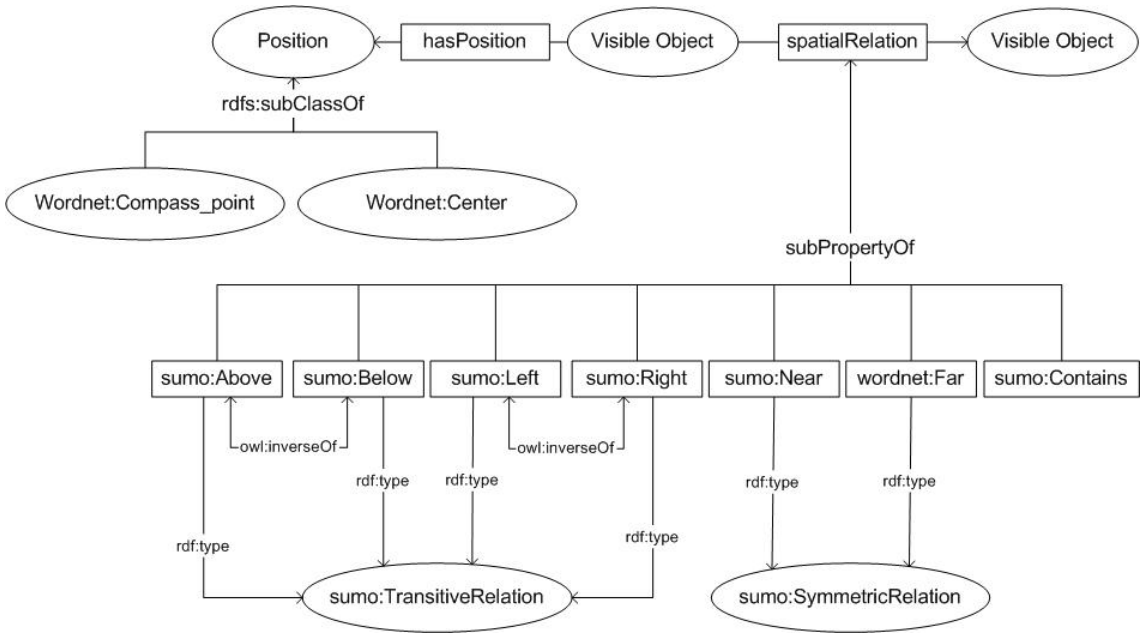

Figure 7.2 Spatial concepts (ellipses) and their properties (rectangles) as they are used in our annotation schema.

the spatial concepts[2]. Figure 7.2 depicts a graph representation of the spatial annotation schema that we use. It shows a `visible object` that has a `position`. The `position` class has two subclasses, namely the WordNet classes compass point and centre. The `visible object` has a spatial relation with another `visible object`. We defined the spatial concepts from SUMO as subproperties of the property `spatial relation`. Left and right are each others inverse, just as above and below. All four are transitive relations. far and near are defined as being symmetric relations. This is a simplification of the concepts far and near as they are used in natural language. Talmy (1983) points out that in human language far and near are not fully symmetric relations: a bike can be near a house, but no one will say that a house is near a bike. This has to do with the size and mobility of the objects, which are properties that we do not take into account in this study.

7.4 Spatial Annotation Tool

The system we propose helps users to add spatial information to annotations of objects in paintings. In order to generate spatial properties of these objects, the image first needs to be segmented into regions. We use a semi-automatic segmentation technique to ensure strong segmentation. First, weak segmentation is done off-line by extracting colour and texture features using Gabor filters. Pixels with similarity values above a given threshold are merged into a region. Each painting in the collection is segmented several times, using different parameters to define scales and

[2]One term from OWL was used, owl:inverseOf, as there is no notion of opposite properties in RDF.

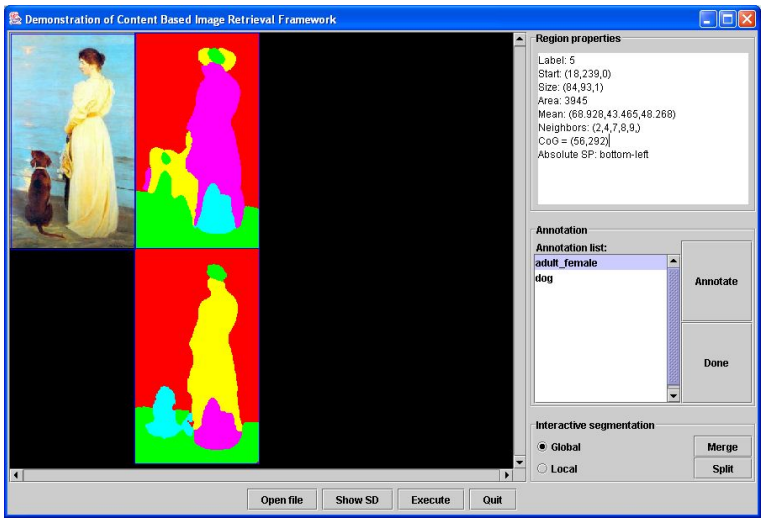

Figure 7.3 Screenshot of the spatial annotation tool, showing a painting segmented at two levels. Region properties of the selected region are shown in the top right corner. Concepts from the annotation are listed in the Annotation list.

thresholds. During the annotation process, the user selects the segmentation that comes closest to a strong segmentation of the image into semantic objects, employing the framework of Nguyen and Worring (2003). The system first offers the user a segmentation of the image using the default set of parameters. The user can now ask for a larger or smaller number of regions, after which the system offers a segmentation with different parameters. This process is repeated until the user is satisfied with the segmentation. By allowing the user to give feedback, the resulting segmented image will closely match the user's expectations.

After segmentation, the user connects concepts in the annotation to regions in the segmented image. This labeling is with done by clicking on a region and clicking on a concept from the annotation. Figure 7.3 shows the interface of the system at the moment that a user is labeling the regions. When the user decides that all relevant regions are labeled, the system continues to compute spatial information of the labeled regions. Absolute positions and spatial relations of the labeled regions are computed. Each labeled region is represented by a bounding box and the centre of the bounding box. Absolute positions of a region are computed by determining in which of nine squares the centre of the bounding box is. For the computation of the spatial relations we employ the method of Abella and Kender (1999). All spatial relations are computed by comparing the centres and borders of bounding boxes of two objects. In Figure 7.4 the definition of left is shown as an example. For details of the other relations we refer to the reference.

Finally, the spatial information is translated into RDF and added to the original annotation, from where it can be queried by other tools. Figure 7.5 depicts a screenshot of the Triple20 toolkit (Wielemaker et al. 2003), that can be used to display and query the annotations. The figure shows the graphical output of Triple20 that displays the spatial annotation of the Matisse painting

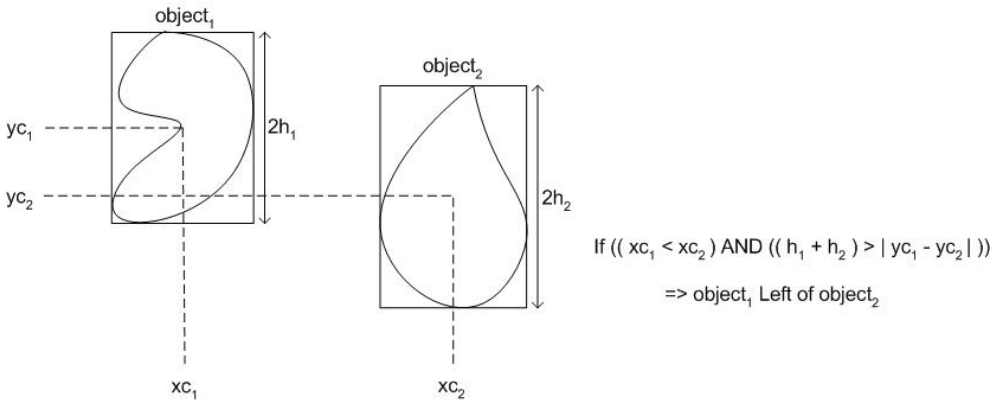

Figure 7.4 Definition of the spatial concept Left.

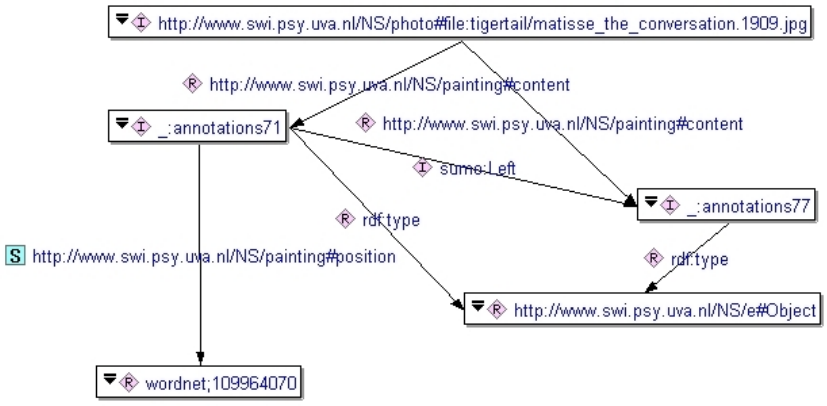

Figure 7.5 Screenshot of Triple20's graphical output of a spatial annotation.

'Conversation' (Figure 7.6) as an RDF graph. The annotation includes two objects linked by the SUMO concept left. The position of one of the objects is specified by a WordNet concept east.

7.5 Preliminary Evaluation

7.5.1 Methods

While designing the tool we have made decisions regarding the types and definition of concepts that were incorporated. We evaluate these decisions by asking two questions:

1. Are the spatial concepts that the tool uses the same as the concepts that users use? In other words, did we choose the right vocabulary?
2. Are the spatial positions and relations that the tool generates for the objects in an image in accordance with positions and relations that users would choose for that image?

Figure 7.6 ‘Conversation’ by Henri Matisse, 1909.

Rashid et al. (1998) asked similar questions for relations between lines and regions. They asked subjects to draw sketches of English-language spatial terms. The sketches were used to map spatial terms onto geometric parameters and their values. One of their results was that topology was more important than metric properties in the selection of spatial terms. We took a different approach: subjects were asked to select spatial terms when provided with a configuration of objects in an image.

For the study we selected eight paintings that were well segmented by the tool (this seems a legitimate criterium since we were not evaluating the segmentation algorithms). Another criterium was that the paintings had to contain at least two objects. Ten PhD students who were familiar with annotation but not in particular with spatial concepts participated in the study. They were split into two groups of five in order to answer the two evaluation questions. We made the assumption that the spatial concepts used by the participants were the correct ones.

Group 1 was provided with the eight paintings annotated with a list of objects that were visible on each painting. They were asked to provide statements about the absolute positions and spatial relations of these objects. Any number of statements was allowed. Comparing the spatial concepts that were used by Group 1 to the concepts included in the tool, will tell us if we chose the right vocabulary and thus provide an answer to Question 1.

Group 2 was also provided with the eight paintings and a list of objects. They were asked to describe positions and spatial relations using a limited list of spatial concepts. The list contained only the terms that were included in the tool. Again, any number of statements was allowed. Comparison of the statements of Group 2 to the statements of the tool provided an answer to Question 2.

7.5.2 Results

Question 1: Vocabulary

In total, 256 statements were written down by Group 1: 132 absolute positions and 124 spatial relations (Table 7.1). Of the absolute positions of Group 1, 81 % were concepts that were included

Table 7.1 Comparison of spatial descriptions by participants to the vocabulary of the tool, divided over absolute positions and spatial relations, in absolute numbers of descriptions and percentages.

Group 1	**Absolute positions**	**Spatial relations**	**Total**
Included in the tool	107 (81 %)	70 (57 %)	177 (69 %)
Not included in the tool	11 (8 %)	36 (29 %)	47 (18 %)
Different level of detail in the tool	14 (11 %)	18 (14 %)	32 (13 %)
Total	132 (100 %)	124 (100 %)	256 (100 %)

in the tool. Only 8 % consisted of concepts that were not included in the tool. This were mainly three-dimensional positions such as ‘in the background’ and ‘in front’. The remaining 11 % of the absolute statements of Group 1 were more precise versions of the concepts in the tool. Examples are ‘almost in the centre’, ‘far right’, ‘between left and centre’.

Of the spatial relations only 57 % of the statements by Group 1 were concepts that were included in the tool, while 29 % of the descriptions were concepts that were not in the tool. The latter were mainly three-dimensional relations (‘behind’, ‘in front of’), statements about the connectedness of two objects (‘connected’, ‘freestanding’) and ‘between’. In addition, participants used concepts that were more precise or less precise versions of concepts in the tool (14 %). ‘Object1 is northwest of Object2’ is more precise than the concepts ‘above’ and ‘left’ in the tool, while ‘Object1 is higher than Object2’ is more general than the concept ‘above’ in the tool.

Question 2: definitions

The five subjects of Group 2 produced a total of 234 statements. Together they selected 127 absolute positions of 27 objects (Table 7.2). Of the 127 positions, 88 (69 %) matched the absolute positions that the tool computed and 39 (31 %) positions did not correspond to the computed positions. This seems a high number of mistakes. However, this could be due to the fact that the tool assigns only one position to each object. If subjects disagree with each other on the position of an object, the tool was not able to match all statements. We found that for only seven of the 27 objects a majority of the participants (at least 3) agreed on a position different from the tool’s position. An example of such a mistake by the tool is the window in the Matisse painting ‘Conversation’. The tool assigned the window the position north, while all subjects agreed that it was in the centre.

Group 2 produced 107 statements about spatial relations. Often, not all possible relations between two objects were described by the subjects. It appeared that they used the inverse Of and symmetric Relation properties implicit: when a subject had stated “woman left of man”, he or she would not also state “man right of woman”. The tool, on the other hand, always explicitly stated the inverse Of and symmetric counterparts of relations. To make the statements of the subjects comparable to the statements of the tool we added symmetric and inverse relations where

Table 7.2 Number of absolute and relative spatial descriptions by participants, divided over descriptions that the tool was able to generate and descriptions that the tool was not able to generate.

Group 2	Absolute positions	Spatial relations	Total
Found by the tool	88 (69 %)	154 (73 %)	242 (72 %)
Not found by the tool	39 (31 %)	56 (27 %)	95 (28%)
Total	127 (100 %)	210 (100 %)	337 (100 %)

necessary. This brought the total number of statements of Group 2 to 210 (and the total number of statements of Group 2 to 337). Of these 210 statements, 154 (73 %) were also found by the tool, 56 (27 %) were not.

Another evaluation measure is the proportion of statements of the tool that corresponds to statements of the subjects. The tool computed 106 statements. A quarter (24) of these statements were about an object pair that was not described by any of the participants, which meant they could not be validated. Of the remaining 82 statements, 56 (68 %) corresponded to at least one participant. Of the 26 'incorrect' statements of the tool, 18 concerned far and near. Participants hardly used these concepts.

7.6 Discussion

In this chapter we explored the possibility of using a content-based image analysis technique to aid the annotation process. We extended manual annotations of objects depicted in an image with automatically derived spatial information. We showed that for this subset of image descriptions, the semantic gap can indeed be bridged. The results of a small evaluation study showed that the generated spatial annotations to a large extent corresponded to spatial annotations by humans: 72 % of all spatial annotations by humans were also generated by the tool. The current vocabulary for absolute positions was sufficient to represent 81 % of the absolute spatial annotations by humans. The vocabulary for spatial relations, however, turned out to be too restricted: only 57 % of the human annotations were accommodated for. The choice of the set of spatial concepts was based on pragmatics, namely those for which automatic detection methods were available. The evaluation showed that this is a severe limitation since people often use three-dimensional concepts, which are very hard to detect. Other frequently used concepts that the tool could not handle were connected and between. Including those concepts would improve the vocabulary of future tools. Another option is to include reasoning to derive spatial information from existing annotations. If an object is, for example, left of another object, and right of a third object, it is likely be 'between' the two objects. Similar rules could be constructed for 'on top of' and 'under'. Two concepts included in the tool were hardly used by human annotators: far and near. It would be interesting to see whether this is also the case in domains other than paintings.

This was an exploratory study with the aim to see whether this approach could work in princi-

ple. Tools such as the E-Culture demonstrator described in Chapter 6 could benefit from automatic annotation with spatial information. It would speed up annotation, in particular for images that depict many objects. We foresee possibilities to automate other parts of the annotation as well. The colour of objects, for example, could be derived using image analysis, and added to the annotation in a similar fashion. In the VisualSEEk system (Smith and Chang 1996), for example, query by sketch is done based on colours and (relative) spatial locations of regions in an image.

The approach described above relies on strong segmentation of the images into objects. If (semi-)automatic segmentation performs badly, manual segmentation can be used to ensure strong segmentation. Ley (2004), for example, uses scalable vector graphics (SVG) to manually define regions and then annotates each region. The E-Culture demonstrator in Chapter 6 also provides a form of manual segmentation, in which rectangles or ovals can be drawn around objects.

8

Building a Visual Ontology for Video Retrieval

In this chapter we set out to combine semantics-based retrieval with content-based retrieval. Existing ontologies, such as the ones described in Chapters 5 and 6 are not easily linked to CBIR techniques as they do not contain visual information about the concepts they describe. We built a 'visual ontology' that contains both general and visual knowledge out of two existing ontologies: WordNet and Mpeg-7. We show how this visual ontology can be used for semi-automatic annotation in a broad domain.

This chapter was co-authored by Marcel Worring and Guus Schreiber, and was published as a short paper in the proceedings of the ACM international conference on Multimedia (Hollink et al. 2005b).

8.1 Introduction

To ensure access to video collections, annotation is becoming more and more important. Ongoing research in video analysis has produced various concept detectors, which are used to detect the presence of a specific concept (e.g. Snoek et al. 2006). This approach to automatic annotation works well within narrow domains, where the number of possible concepts is small. However, this becomes difficult as soon as the collection gets broader. Domains like biology, art, family pictures and broadcast news are problematic, as it is infeasible to build detectors for all possible concepts.

One approach to this problem is to use background knowledge about the domain under consideration. There is structured background knowledge available about various topics, in the form of ontologies or thesauri. Examples are SnoMed, MeSH and the Gene Ontology for health care and AAT and Iconclass for art. Also, non-domain-specific ontologies exist, such as WordNet and Cyc. Ontologies are used in annotations for various reasons. If existing, well-established ontologies are used, they provide a shared vocabulary. Not only the terms themselves are agreed upon, but also the meaning of the terms, since the meaning is partially captured in the (hierarchical) structure of the ontology. Polysemous terms can be disambiguated and relations between concepts in the ontology can be used to support the annotation and search process (Chapters 5 and 6). Ontologies are currently used for manual annotation (Hyvönen et al. 2004b, Schreiber et al. 2001). They are, however, not suitable for automatic annotation based on the visual properties of an image, since they contain little visual information about the concepts they describe.

In this chapter we aim to build an ontology that does contain visual information about general, high-level concepts, and investigate the use of such a 'visual ontology' for semi-automatic annotation in a broad domain. We identify the requirements of a visual ontology. Following these requirements, we build a visual ontology out of two existing ontologies: WordNet and Mpeg-7. WordNet contains visible concepts, such as material and colour, but this visual information is not structured in a way that makes it useable for annotation. The visual concepts are not linked to general concepts; there are no statements saying that a boat is capable of motion or that houses are made of brick or wood. The Mpeg-7 ontology (Hunter 2001) consists of visual properties and classes, but does not contain general, high-level concepts. In the presented visual ontology, we create the links between visual properties and general concepts that are lacking in existing ontologies.

The aim of the visual ontology is to facilitate quick semantic annotation. This can be done by detecting visual properties in a video and then searching the visual ontology for concepts with matching properties. In this way, the list of possible annotations is reduced to a manageable size from which relevant annotation concepts can be selected. In addition, the visual ontology can be used to support search. Search with the visual ontology can be done by looking for regions within a video whose visual characteristics match the visual properties of a query concept taken from the visual ontology. This will reduce the video collection to a smaller set of shots, through which the searcher can browse to find relevant items.

We explore how annotation with the visual ontology performs under ideal circumstances. We evaluate annotations of 40 shots of news video made with the visual ontology, and discuss the added value of each visual property in the ontology.

This chapter is organised as follows. Section 8.2 contains related work. In Section 8.3 we identify requirements for a visual ontology and in section 8.4 we present the design of the visual ontology based on these requirements. We discuss choices we made regarding the use of existing ontologies, the incorporation of visual properties and the population of the visual ontology with instances. Section 8.5 contains the setup and results of a small evaluation study. Finally, section 8.6 contains a discussion.

8.2 Related Work

Research has been done on using ontologies for retrieval of textual resources (Handschuh and Staab 2003a). However, using ontologies for retrieval of visual resources is a relatively new area. Hauptmann (2004) proposed to design an ontology of automatically detectable concepts that could provide a basis for annotation of broadcast video. Ongoing research will tell which concept detectors are suitable for inclusion in such an ontology.

Mezaris et al. (2004) combine a thesaurus with relevance feedback. They let users describe high-level keywords with terms from a small ontology of visual descriptors such as luminance and size. Descriptions of the keywords are compared to extracted features of regions in video footage and matching regions are returned. Relevance feedback is then used to refine the result set. Bertini

et al. (2005) propose an ontology that contains not only words, but also images to capture concepts that are difficult to express with linguistic concepts.

Tansley et al. (2000) investigate the automatic building of a multimedia thesaurus. They use existing annotations of images to connect example images to high-level concepts. They extract low-level features from the example images and assign these to the high-level concepts. The low-level features can then be used to classify images without metadata. Bloehdorn et al. (2005) present a system with a similar aim, namely the automatic linking of visual features to domain concepts. They present a tool for manual semantic annotation. While annotating, the user is at the same time creating an example set of images. The system extracts visual features from the example images and links these to annotation concepts. They use several ontologies including a visual descriptor ontology (based on Mpeg-7), an ontology of the domain under consideration and a core ontology (Dolce). We take a step back from Tansley et al. and Bloehdorn et al. and investigate the desired characteristics of a visual ontology and how it can be used, rather then how it can be automatically created.

Hoogs et al. (2003) and Stein et al. (2003) manually extend WordNet with visual tags describing visibility, different aspects of motion, inside/outside and frequency of occurrence. They describe a system that uses the extended WordNet for annotation of video sequences. The system firstly analyses the videos to detect general low-level features such as colour and motion, as well as a limited number of high-level concepts. Secondly, both the low-level features and the high-level concepts are used to search the extended WordNet for relevant annotations. In addition, Hoogs et al. use the detected concepts to search the WordNet glossary. The glossary is a collection of natural language descriptions of concepts. Even though searching the glossary may lead to good results for some searches, the performance is unpredictable, since the structured knowledge of the ontology is no longer used. The papers discussed above do not explicate their rationale for including certain visual characteristics in their visual ontology, nor do they clarify why they chose a particular representation and format for their ontology. In addition, it is not clear how much of the retrieval performance of the systems can be attributed to the visual ontologies, and how much is caused by the quality of the detectors. In this chapter we intend to address these issues. Our work builds on the work of Hoogs et al. and Stein et al. in the sense that we further investigate the idea of using WordNet as a starting point for a visual ontology. We follow Bloehdorn et al. in the sense that we aim to use the visual ontology for semantic annotation, and that we use Mpeg-7 to represent the visual characteristics.

8.3 Requirements

A visual ontology needs to contain classes and properties that describe visual information, such as colour and shape of objects. These properties need to be *visually perceptible*; characteristics like mass, smell and status are thus disregarded. In order to support automatic annotation of visual resources, the property needs to be *detectable* from the visual data. Finally, visual properties need to be *discriminating*; in at least some cases the value of a visual property needs to tell something

about the class an object belongs to. If too many objects have the same value for a visual property, it is not discriminating and therefore not useful for a visual ontology. For our purposes of annotation in a broad domain, a visual ontology also needs to comprise terms from a broad domain. It needs to contain relations between the visual and general concepts, such as the statement that a wheel is round.

Since we aim to be application independent, the ontology as well as the resulting annotations need to be usable and reusable by various applications. Using existing knowledge corpora and standards instead of introducing our own vocabulary increases interoperability. The need for interoperability of data has been widely recognised in the semantic web and digital library communities (see e.g. Lee 2004). Martínez et al. state that ”one of the current trends is that the content is created only once, but it should be accessible via any access network and client device” (Martínez et al. 2002, p.1).

In summary, a visual ontology needs to:

1. contain visual concepts, that are both detectable and discriminating.
2. contain general concepts from a broad domain.
3. contain relations between general and visual concepts.
4. be interoperable.

Hoogs et al. (2003) and Stein et al. (2003) extended WordNet with visual terms. The use of WordNet ensured that they met the generality requirement, while the added visual properties fulfilled the requirement for general-visual relations. We seek to build an ontology that meets all four requirements.

8.4 Design of a Visual Ontology

8.4.1 Using existing ontologies

We choose Mpeg-7 to describe the visual information. Mpeg-7 is a standard for describing multimedia content published by the Moving Picture Experts Group (Mpeg) (Martínez 2001). It is aimed at a broad range of applications. The Mpeg-7 OWL ontology as published by Hunter (2001) contains low-level visual properties like colour, shape and motion.

Following the choice of Hoogs and Stein, we used WordNet as a general ontology. WordNet is a widely used lexical database in which nouns, verbs, adjectives and adverbs are organised into synonym sets (synsets), each representing one underlying lexical concept (Fellbaum 1998). Containing over 100,000 synsets, its broadness makes it suitable for annotation in broad domains. We used the RDF/OWL translation of WordNet by van Assem et al. (2004).

The broadness of WordNet combined with the visual information in Mpeg-7 ensure compliance with the first two requirements of generality and visuality. By using an ISO standard like Mpeg-7 and a widely used lexicon like WordNet we seek to fulfill the fourth requirement of interoperability. Furthermore, by using RDF/OWL as the language for our visual ontology, we ensure

interoperability also on the syntactic level. In order to link general concepts to visual concepts and meet the third requirement, we add statements of the form:

general concept	-	**visual property**	-	**visual value**
wordnet:subway	-	VO:environment	-	wordnet:indoor
wordnet:car	-	mpeg7:motion	-	wordnet:rigid.

8.4.2 Selecting visual properties

As identified in Section 8.3, a visual property needs to be visually perceptible, detectible and discriminating to be useful in a visual ontology. We use two sources of visual properties. Properties from the Mpeg-7 ontology (Hunter 2001) were used for the most low-level descriptors. From 'Mpeg-7 Visual' we used colour, shape and motion. Mpeg-7 provides one more descriptor that is both visual, detectable and discriminating: texture. We did not incorporate this directly in our visual ontology because texture is the main feature to detect `materials`, which was incorporated as a property in the visual ontology. Geusebroek and Smeulders (2005), for example, used texture to detect materials.

While Mpeg-7 offers low-level visual properties, WordNet contains more high-level visual properties. The hierarchy under property/attribute contains the following concepts that meet the requirements for visual properties: visibility, naturalness, environment and material. Visibility can have the values invisible or visible. Visible can be further refined into visualisable and viewable. A visualisable concept was defined by Stein et al. as a concept that "one can not only see [..], but also draw" (Stein et al. 2003). All instances of a visualisable class must have visual characteristics in common like shape and colour. We consider a concept viewable if it can be seen in today's daily life without instruments. Microbacteria, intestines and galleons are not viewable.

The RDFS graph of the visual ontology is depicted in Figure 8.1, showing that there are seven visual properties that link the top-level WordNet concept entity to visual values from WordNet and Mpeg-7. Subclass relations are denoted by arrows, property names are in italics and class names are in normal font. The values that the visual properties can take were all concepts from WordNet. The set of possible values is large for some properties, such as for `material`, which can have values from all WordNet subclasses of cloth, building material and matter. In this way we avoided restricting the usability of the ontology to a particular domain or use case. Also, an annotation is never too specific, as in an ontology a specific concept can always be traced back to its more general parent. The classes motion, colour and shape are modeled as subclasses of corresponding Mpeg-7 classes, to avoid changing the the Mpeg-7 classes themselves (Assem, van et al. 2004).

8.4.3 Populating the visual ontology with instances

As a proof of concept, we implemented part of the visual ontology: all classes in the WordNet hierarchy under the class conveyance were extended with visual properties using the schema described above. The hierarchy under conveyance contains 564 classes. This number makes it feasible to

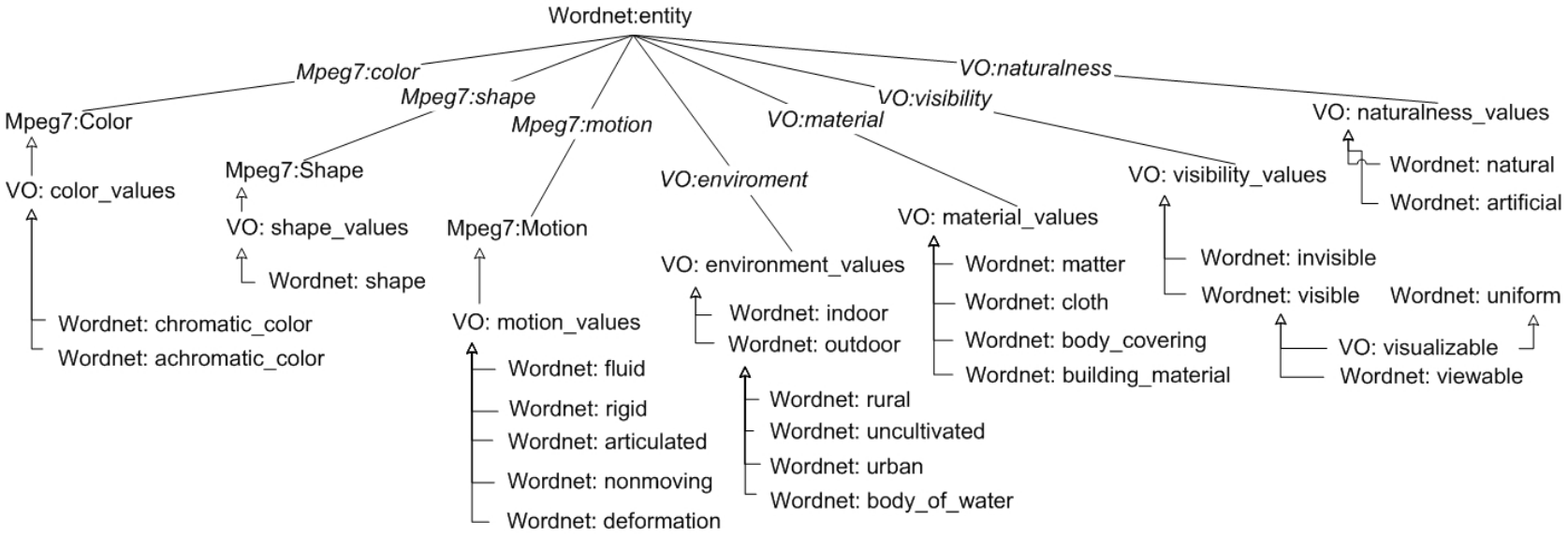

Figure 8.1 RDFS graph of the visual properties from Mpeg-7 and WordNet and their values from WordNet.

manually assign the values of the visual properties, while it is still large enough to demonstrate annotation in a broad domain. We made use of the hyponym relations in WordNet; if a class had a certain value for a property, all its hyponyms were assigned the same value.

Many classes in WordNet can have more than one value for a property. A bicycle, for example, can be seen in both urban and rural environments. In these cases, we assigned both values. Only visual values that are likely to be encountered were assigned to a concept; a bicycle can in theory also be found indoors, but we did not include this unlikely value in the visual ontology. In many cases no value could be assigned to a property. An example is the shape property of the class public transport, since this class has too many possible shapes. In total, 548 property-value pairs were assigned, excluding the pairs that were deduced from the hyponym hierarchy.

8.5 A Thought Experiment

8.5.1 Setup

An evaluation of the *search* capabilities of the visual ontology would require the detection of visual characteristics of all video's in a collection, which is beyond the scope of this chapter. Instead, we evaluate the quality of *annotations* made with the visual ontology. To decouple detector quality and the power of the visual ontology, we make the assumption that we have perfect detectors for the six detectable properties and their categorisation into symbolic values. Currently, detectors exist for each property and detector quality is improving fast, as can be seen from Snoek et al. (2004). This makes it plausible that good detectors will be available in the near future.

40 Shots from the TRECVID 2003 collection that contain a form of conveyance were selected for testing. The shots display boats, trains, cars and planes, with 10 shots in each category. Firstly, keyframes of all shots were segmented (Hoang et al. 2002). Secondly, property values were assigned manually to the region in each keyframe that depicts a form of conveyance. The value for the 'environment' property was determined by looking at the values of the neighbouring regions. Finally, the visual ontology was searched for viewable concepts matching the list of property values of a region. In the case of conveyance, almost all concepts are viewable. The resulting concepts were returned as possible annotations of the video.

8.5.2 Results

We evaluated the results with two measures, precision and reliability. Precision is defined as the number of relevant annotations found, divided by the total number of annotations found. Reliability is the percentage of shots for which at least one relevant annotation is found. Relevance of an annotation was decided manually, based on the free text descriptions (glosses) that WordNet provides for all concepts. Figure 8.2 shows examples of retrieved annotations. At least one relevant annotation was found for 93 % of all shots (Table 8.1). Precision of the annotations has a mean of 8 %. On average, the 564 concepts of the conveyance hierarchy have been reduced to 57. This is a manageable list from which relevant annotations can be picked manually or automatically.

Correct – taxicab

Incorrect descriptions electric-car, berlin, limousine, minicab, gypsy cab.

passenger train, commuter train

freight train, bullet train, tandem trailer, articulated lorry, trailer-truck, helicopter, sky-hook, freight-liner, ladder truck, single rotor helicopter, shuttle helicopter, cargo helicopter.

Figure 8.2 Correct and incorrect annotations selected from the visual ontology based on the visual characteristics of the shots.

Table 8.1 Mean number of generated annotations, correctly generated annotations, precision and reliability.

Shot depicting	Trains	Cars	Boats	Planes	Total
Retrieved annotations	61.0	37.7	37.8	91.0	56.9
Correctly retrieved annotations	1.9	3.8	2.3	9.2	4.3
Precision	0.03	0.10	0.06	0.10	0.08
Reliability	0.90	1.00	0.80	1.00	0.93

Retrieval relied mostly on only three visual properties: material, motion and environment. Colour, shape, visibility and naturalness appeared to be less useful. Colour and shape were detected in the shots, but the visual ontology contained few concepts that had a value for colour or shape. This is because the majority of the WordNet concepts are not visualisable and do not have a fixed shape or colour for every instance. In the domain of conveyance, almost all concepts are viewable, with the exception of some historic or fantastic vehicles like galleon, which are visible but not viewable. All WordNet classes in this domain have the value artificial, except for pirogue, which is a canoe made out of a whole tree.

8.6 Discussion

In this chapter we described our experiences with the creation and use of a visual ontology in a video collection. We propose that a visual ontology for this purpose needs to contain broad general concepts, visual descriptions and links between those two. Furthermore it needs to comply with existing standards. A combination of two existing ontologies, Mpeg-7 and WordNet, meets these requirements. Visual properties that are represented in the visual ontology need to be visually perceptible, detectable and discriminating. In a study in the conveyance domain, it appeared that the visual properties `visibility` and `naturalness` were not discriminating in this domain: too many classes had the same value for these properties. We believe, however, that in other domains these properties could be important discriminating attributes. `Visibility` can be expected to play a role in other subdomains of broadcast news, such as politics, while `naturalness` is important for domains in which, for example, landscapes are depicted. Therefore, we did not remove them from the visual ontology.

Colour and shape were not useful as visual properties, since most of the WordNet conveyance concepts cannot be described in terms of colour and shape. In other words, conveyance concepts are not visualisable. Further study is needed to determine whether colour and shape are useful in other domains, such as botanical or food domains. In this study, motion, material and environment were both detectable and discriminating and therefore the most important properties for retrieving relevant annotations.

Using existing knowledge bases ensures interoperability. However, it also means that one has to use resources with a design that may not be ideal for the current purpose (see also Chapter 5). WordNet is not ideal for visual descriptions, since the hierarchy is more functionally orientated then visually, as was pointed out by Stein et al. (2003). This means that all members of a class have functional properties in common, but not necessarily visual properties. In addition, WordNet is originally not a subclass hierarchy, but a hierarchy of hyponyms. Synsets represent lexical concepts. As language is not always consistent, WordNet is not always consistent. Because of this, visual properties of a concept do not always propagate to all its hyponyms, which complicates the process of assigning visual properties to classes.

One of the questions we asked was ”How well can a visual ontology perform in ideal circumstances?” We successfully reduced the list of possible annotations to a list of manageable size

from which relevant annotations can be picked. Going through the result can be done by a human annotator. Another option is to combine the results of the visual ontology with text associated with a shot. After detecting the visual properties of a region, a visual ontology can tell which of the words in the textual description have visual properties that correspond to the visual properties of the shot.

Although the reliability of the approach was high (i.e. for the majority images at least one correct annotation was found), we should note that in some cases the list of possible annotations was too restricted. The list of possible annotations for the train in Figure 8.2, for example, did not include the class public transport. This is an artifact of our retrieval method, which only returns classes that have the correct value for all visual properties that were detected in the shot. Metal was detected in the shot, and since the class public transport does not have a value for material, it is not returned. This strict use of the visual ontology restricts the lists of possible annotations too much. The opposite approach, in which all classes are returned unless it is known that they cannot have the detected visual properties, would result in very long lists of possible annotations. A possible solution would be to introduce a ranking in the annotation list. Concepts matching all detected visual properties appear highest in the list, while concepts with unknown values for some of the detected concepts appear lower. The list of possible annotations for the train in Figure 8.2 would then include train at a high rank and public transport at a low rank.

Extending the visual ontology from only the hierarchy under conveyance to the complete WordNet hierarchy will increase the size of the result list: it will occur more often that concepts in different places in the hierarchy have the same set of property values. One way to overcome this is to use more visual properties in the visual ontology, as the range of available detectors increases. Hoogs et al. (2003) deal with this problem by employing a limited set of high-level concept detectors as a starting point for a search through their visually extended WordNet. In this way classes in different parts of the hierarchy with equal visual properties can be disambiguated. Detectors for this purpose should not be too specific, as they are meant to start a search rather than find a specific concept. In addition, they should correspond to one or more concepts in WordNet. Wordnet concepts that meet these criteria and for which detectors are available (see e.g. Snoek et al. 2004) are, for example, animal, human and vegetation.

Acknowledgements

We would like to thank Niek Fraanje for his valuable contributions to the requirements analysis.

9

Conclusions and Discussion

Retrieval of images and video poses specific challenges within the field of information retrieval. This is due both to the many possible descriptions and interpretations of one visual resource and to the semantic gap between what can be automatically derived from the raw data of a visual resource on the one hand and the human interpretation of a visual resource on the other hand. The focus of this thesis is on the process of retrieval of visual resources, including user information needs, query formulation, annotation and search. We study problems related to visual-resource retrieval and develop solutions that are based on extending the semantics of visual-resource descriptions through both background knowledge and image-analysis techniques.

The main contributions of the thesis are:

- A framework for descriptions of visual resources, which showed its value in three different application contexts.
- Methods and application scenarios to support annotation and search with structured background knowledge.
- Insights into the nature of the semantic gap in different domains, ranging from narrow to broad.
- Methods to combine structured background knowledge and image analysis.

In this chapter we revisit the four questions raised in Chapter 1, discussing methods used and results achieved. The final section reflects on the work and discusses future research.

9.1 Research Questions Revisited

9.1.1 How do people describe and search for visual information?

A prerequisite for creating richer descriptions of visual resources is to know the spectrum of descriptions that can be constructed for a visual resource, and the types of descriptions that are used in practice. To this end, a framework was built for the classification of visual-resource descriptions (Chapter 2). The framework classifies queries as well as annotations: both are considered to be special cases of descriptions. A number of classification methods exist in literature, each with its

own focus and viewpoint on visual resources. Several of these well-known methods and standards in literature were combined in the framework, including the VRA Core Categories, the Panofsky/Shatford model, the ten-level model for indexing of Jaimes and Chang and Jörgensen's empirically found classes of image attributes. The framework consists of classes and class-relations that are derived from the categories, facets, levels and classes that constitute these methods in the literature.

The classification framework distinguishes three viewpoints on images, namely the non-visual metadata level, the perceptual level and the conceptual level. For every viewpoint a set of descriptive classes and class-relations are specified. The information at the non-visual level is about the context of the resource rather than about the content. The classes at the non-visual level are a subset of the VRA element set. The perceptual level comprises descriptions that are directly derived from the visual characteristics of the resource. No deep knowledge of the world or the domain is required at this level. The conceptual level gives information about the semantic content of the image. World knowledge is required for descriptions at this level. The conceptual level is further divided into three sublevels: general descriptions that require only everyday knowledge of the world, specific descriptions in which the objects and scenes are named and for which domain-specific knowledge is required and abstract descriptions that are interpretative and subjective.

The framework does not commit to a particular application, domain or query language. This unbiasedness makes it possible to use parts of the framework in different contexts. In this thesis the framework was used to classify queries and topics in Chapter 4 and also to construct metadata schemas for cultural heritage domains in Chapters 5 and 6. The strong link with existing literature makes it possible to compare descriptions classified in the framework to descriptions structured with existing methods.

In order to determine which categories of the framework are used most frequently for description of visual resources, we used the framework in three different application contexts. The use of categories in the framework can vary across domains and retrieval methods. Therefore, we chose three contexts that are increasingly domain- and retrieval method-specific: a setting not related to a domain or method (Chapter 2), the domain of paintings (Chapter 7) and the domain of broadcast news for a content-based image retrieval system (Chapter 4).

The first setting was as generic and unbiased as possible with respect to domain and retrieval method (Chapter 2). It was found that people performing a 'category search' task use object descriptions twice as much as scene descriptions. General descriptions were by far the most frequently used level (74 %), followed by specific descriptions (16 %) and abstract descriptions (9 %). Frequently used classes were events, places and spatial descriptions. People used more specific and less abstract descriptions in 'search tasks' than in 'describing tasks'. Considering the infrequent use of abstract descriptions and the fact that this is also the level that suffers most from interpretability as explained in Chapter 1, we regard retrieval based on abstract descriptions as infeasible for many domains.

The class of spatial descriptions was studied in detail in the second setting. We found that people describing paintings use absolute positions of objects as much as relative positions, including

three-dimensional relations and notions of connectedness.

In the third setting, descriptions of information need (topics) and textual user queries were studied in the domain of broadcast news on the MediaMill 2003 system, which is an interactive news video retrieval system based on automatic indexing and query-by-example (Chapter 4). The majority of the queries was about general objects. Specific queries were more frequently used than in the first context and led to better results than general queries.

Overall, we see that the specific level is more important in the news domain than in the other domains. Also, 'search tasks' result in more specific terms than 'describing tasks'. These observations are consistent with results of Armitage and Enser, who studied queries on several 'news like' collections. Search tasks also contained less abstract descriptions than describing tasks, which is a confirmation of findings of Jörgensen.

Information about the structure and components of user queries can help to improve query and annotation interfaces. It can be used to determine what types of descriptions should be facilitated and what classes should be part of the vocabulary. We used the categories of the framework plus the information about the frequency of use of the categories as the basis for an metadata schema to structure descriptions in the cultural-heritage domain of paintings (Chapters 5 and 6). Slots in the schema correspond to categories of the framework. The values of the slots are limited to relevant parts of ontologies, thus further structuring and constraining annotations. The metadata schema covers descriptions of both art-historic information and the content of an image. The slots for art-historic information are VRA elements, as VRA provides an adequate covering of this type of annotation. For content slots, however, VRA is not sufficient. Paintings require a finer-grained description of content than provided by standards such as Dublin Core and VRA, which contain just a single Subject element. Therefore, we extended the VRA Subject element with additional slots, modeling them as a specialisation of the Subject element, so that they can be 'dumbed down' to the original VRA standard.

The metadata schema was used in two annotation and search systems, the Mia demonstrator (Chapter 5) and the E-Culture web demonstrator (Chapter 6). The first provides a schema for content annotation that allows constructing content descriptions of the form 'agent-action-object-recipient'. This relatively complex structure was designed to resemble a natural-language description. However, it also restricted the descriptions to fit this particular format. The subsequent content-annotation schema of the E-Culture web demonstrator was less structured and with that applicable to a broader range of descriptions. It allows content descriptions of the form 'subject-relation-object'. The E-Culture demonstrator provides an interface through which the annotator can indicate image objects and object relations, which are named either with concepts from an ontology or with literals. The relation can either be a verb, making the description resemble a simple natural-language sentence, or a relation between the subject and object. The use of the metadata schema in the Mia and E-Culture demonstrators enabled us, for example, to distinguish paintings made in Paris from paintings depicting Paris and to distinguish paintings of a woman holding a hat next to a man from paintings of a woman holding a man.

The repeated use of the framework within this thesis, the fact that it integrates well-known

literature and the fact that no omissions or errors were found indicates that the framework is useful for classification of descriptions of visual resources. The framework has shown its value for comparing the use of categories among domains. A real proof of the added value can only be obtained through heavy and continued use by real users or by comparison with similar frameworks, neither of which are currently available.

9.1.2 How can structured background knowledge about the domain be used to support the process of visual-information retrieval?

Thesauri have been used as controlled vocabularies in cultural heritage institutions for several purposes. The use of widely-known vocabularies ensures that annotations and queries can be understood by everyone who is aware of these vocabularies. In addition, the thesauri aid users in finding the right terms and make sure that all users use the same terms. Our hypothesis was that in knowledge rich domains, in which large bodies of structured background knowledge are available and experts agree on the main concepts and relations, ontologies not only serve as structured vocabularies, but also support disambiguation, speed up annotation and improve recall and precision.

To support this hypothesis we have developed application scenarios in which multiple ontologies together form the vocabulary (Chapters 5 and 6). Links between slots in the metadata schema and relevant parts of the ontologies restrict the use of the vocabulary. This restriction makes it easier for a user to find the right concept for annotation or search. Values of the creator slot, for example, are limited to concepts from ULAN, a thesaurus that contains artist names.

Disambiguation of homonymous terms is facilitated by showing the parts of the hierarchy containing the term in a browser window. This way, disambiguation is made easy for the user as the position of a word in the hierarchy clarifies its meaning. Disambiguation is especially important if multiple large vocabularies are used, which is the case in both Chapters 5 and 6.

Ontologies speed up annotation when relations between concepts in the ontology are used to derive values of slots in the metadata schema. We have shown in Chapter 6, by means of a use case, that from the value of the creator slot, annotations can be derived of the material, culture and style of the painting. Extending the annotation by deriving values will quicken the completion of an annotation and thus lighten the burden on the annotator.

Similarly, relations between concepts in the vocabulary can be used to extend a query and with that improve recall (Chapter 5). A query for Flower will return paintings annotated with Sunflower since there is a hyponym relation between sunflower and flower. A search for cubist paintings will return paintings annotated with Picasso as the creator if there are links between creators and styles. A query for Venus will return paintings depicting Aphrodite, since the ontologies contain an equivalence relation between them. However, too much expansion of the query may lead to low precision. In an experiment using all WordNet relations, we identified patterns of relations that increase recall while preserving precision (Chapter 6). Expansion with hyponyms (cf. the flower - sunflower example) is intuitive, and the experiment confirmed that this improves recall

while maintaining precision. We found that retrieval results can be further improved if other types of relations are used as well. Query expansion with a combination of hyponym, holonym and meronym relations provides a good balance between recall and precision. Hypernym relations also improve search results, but a combination of hypernym relations with other types of relations (e.g. hyponym or holonym) lowers precision more than it increases recall. More then three intermediate nodes between the query concept and the annotation concept appeared to harm precision. We can conclude that semantic annotation and search systems such as the ones described in Chapters 5 and 6 can improve their recall values by expanding query results with not only hyponym relations, but also with part-of and hypernym relations.

9.1.3 What are the circumstances under which the semantic gap limits retrieval?

We addressed the influence of the semantic gap in the three domains: organic cells, paintings and broadcast news. The domain of organic cells is visually the simplest of the three. For some parts of the cell domain, a direct link from visual features to high-level concepts could be made, while for other parts this proved infeasible. We concluded (Chapter 3) that this domain is on the borderline of where retrieval is limited by the semantic gap. We performed a study that allowed us to explicate a number of key domain requirements for the ability to have direct links between results of image analysis and domain concepts. The first requirement is that strong segmentation needs to be possible; if retrieval requires the identification of objects (this is not necessarily the case as there are approaches that use visual features of the image as a whole), then these objects need to be well-discernable. The second requirement is that the domain is narrow, well-understood and concepts are widely agreed upon so that subjectivity is minimised. Thirdly, classes need to have small visual variance and be clearly visibly distinguishable from other classes. In practice, creating a model that links visual features to domain concepts requires familiarity with image analysis, which domain experts usually do not have. Therefore, an additional requirement is that the model is made either automatically (e.g. by machine learning) or as a joint effort of domain experts and image analysis experts.

The domain of paintings does not fulfil all these requirements, but a subset of painting descriptions, namely spatial descriptions of image objects, proved to be an area in which the semantic gap is small (Chapter 7). As a first step, the image was segmented into regions representing objects. To ensure that the requirement of strong segmentation was met, a semi-automatic segmentation technique was used. The objects were then manually annotated with concepts from an ontology. Subsequently, the positions of objects within the painting and the spatial relations between them were extracted from the visual data of the image and added to the annotation.

Automated retrieval is known to be difficult for broadcast news, due to the broadness of the domain. Also with respect to other requirements the news domain scores low. Firstly, segmentation of news images is difficult. Secondly, the domain is not well understood and agreed upon in the sense that it changes continuously, it is prone to different interpretations and it contains a wide range of objects, persons and events. Thirdly, there is large visual variance within classes and little

visual distinction between classes. A study on the content-based MediaMill 2003 system (Chapter 4) confirmed the negative effect of a wide semantic gap on news retrieval. The study revealed a number of additional issues that are related to the semantic gap. We found no correlation between a user's familiarity with a topic and quality of search results. This finding indicates a discrepancy between a user's interpretation of the topic and what is retrieved based on visual features. A second observation was that the use of speech transcriptions contributed more to retrieval than the use of image analysis. The effect was stronger for 'specific' topics than for 'general' topics. This suggests that different strategies should be applied to different topic types: emphasis should be on text for specific topics, while it can be on both text and low-level visual features for general topics. Future retrieval systems could benefit from a (automatic or manual) classification of the topics, in order to adapt the retrieval strategy.

9.1.4 How can structured background knowledge and image analysis be combined to improve visual-information retrieval?

A solution to reduce the problems caused by the semantic gap and at the same time reduce the time and effort of the annotator, was found in a combination of image analysis techniques and techniques using structured background knowledge. The combination was first made by means of semantic web rules in the domain of organic cells (Chapter 3). RuleML rules linked visual features of cell images to concepts from a medical ontology. The rules modeled a part of the domain knowledge that was not present in the domain ontology, namely the visual appearance of concepts. This made it possible to automatically annotate images based on their visual features.

The combination was also made in a subset of the news domain, in which the visual appearance of concepts is less straightforward (see Section 9.1.3). Concepts in WordNet were enriched with visual properties and values (Chapter 7). The resulting 'visual ontology' can be used in annotation and search. For annotation, visual properties of a region in a news video are detected, after which the visual ontology is searched for concepts with matching properties. This results in a list of possible annotations of the region, from which relevant annotations can be selected. In this way, the amount of possible annotations is reduced to a manageable number. Search with the visual ontology can be done by retrieving visual resources whose visual characteristics match the visual properties of the query concept. This will reduce the collection to a smaller result set, through which the searcher can browse to find relevant shots.

For paintings, this approach is not feasible since objects in a painting do not have a sufficiently fixed appearance. Instead, image analysis and background knowledge were used side-by-side, complementing each other: those parts of the annotation that are suitable for image analysis (i.e. parts with the characteristics listed in Section 9.1.3) are provided automatically, while parts of the annotation that require a level of semantics or precision that is not provided by image analysis, are filled in by a human annotator, if possible with help of background knowledge as discussed in Section 9.1.2. Using this approach, manual annotations of objects in paintings were augmented with automatically derived spatial properties of the objects (Chapter 7). The spatial information

was added to the annotation.

9.2 Discussion and further research

Three domains were used to answer the research questions: organic cells, paintings and broadcast news. The differences between the domains are both an asset and a drawback. On the one hand, the differences make comparison of results difficult. The same precision and recall numbers, for example, can be impressive in one domain while mediocre in another. On the other hand, different domains were necessary for the exploration of different techniques. The characteristics of a domain for a large part determine the applicability of a retrieval approach: knowledge-rich domains benefit from the application of ontologies, domains with a clear link between low- and high-level concepts are suitable for image analysis and domains that have neither sufficiently elaborate bodies of structured background knowledge, nor a clear enough visual link between low-level features and high-level concepts, require a combination of techniques. The news domain, for example, could benefit from a combined approach, since image analysis alone does not give satisfactory results and the background knowledge available is fairly general.

We have shown that the domain of paintings is knowledge-rich and benefits from structured background knowledge. Annotation vocabularies range from narrow to broad and from unstructured to highly structured. There is a trade-off between expressiveness and ease of sharing. On the one hand, small vocabularies aid retrieval by ensuring correspondence between annotations and queries. Similarly, highly structured metadata schemas are an asset, since they restrict the interpretation of concepts in the annotation or query. On the other hand, the vocabulary and metadata schema should be sufficiently expressive to facilitate a wide range of descriptions, or a second type of semantic gap will emerge: the discrepancy between what can be said about an image with a given vocabulary and the interpretation that a user has of that image. Further research is needed to find the optimal balance between expressivity and structure. The content-annotation schema in Chapter 6 was designed to provide more expressivity and less structure than the schema used in Chapter 5.

The research in this thesis has for a large part focussed on speeding up, enhancing or enriching manual annotation. Manual annotation is sometimes considered infeasible, because it is laborious (Hyvönen et al. 2004a), time consuming (Little and Hunter 2004) or costly (Smith et al. 2005). However, there are strong indications that manual annotation will continue to play an important role in visual-resource retrieval. Annotation practices in cultural heritage institutions today indicate how much time and resources people are willing to spend on correct and elaborate descriptions of the items in their collections. Another indication is the eagerness that people and organisations exhibit to share information on the web. The amount of HTML pages written with the purpose of sharing text makes it plausible that a similar effort will be made to share images. Especially if semi-automatic tools emerge that facilitate quick and intelligent annotation, people will be prepared to spend time and resources in order to produce large numbers of correct, high-level annotations of visual resources. The work in this thesis has been a step towards the design of

such tools.

We have made a point of incorporating an evaluation or user study in every step of the research. However, this thesis does not contain a user study with real users, who are annotating or searching as part of their daily workflow. The reason is that the behaviour of real users is necessarily linked to the characteristics of their collection, the limitations of the vocabulary or the peculiarities of their retrieval system. In order to conclude on the way people describe images without the bias of a collection, vocabulary or retrieval system, we invented an artificial but realistic retrieval task in Chapter 2. More general statements about the way people describe images would require studies in a larger number of contexts than the three contexts explored in this thesis. Generally speaking, more evaluation on real life situations is needed to evaluate new methods not only based on the number of correct annotations they produce, but also on how much they contribute to the goal of retrieval and the added value over current retrieval practices.

Content-based image retrieval (CBIR) can be used for automatic annotation in domains in which there is a direct link from low-level features to high-level concepts. In more complicated domains, CBIR has proven its value in combination with other techniques such as text retrieval and natural language processing. In this thesis we have combined CBIR with ontology-based retrieval. We have shown that this combination has the potential to produce semantic annotations in an efficient manner. Extensions of our work are possible in several directions. Structured background knowledge can be incorporated in CBIR systems such as the one described in Chapter 4; knowledge about classes and their instances can be used to solve general queries by searching for specific instances. The finding in Chapter 4 that specific queries are better retrieved than general queries, suggests that this will increase performance. The same principle can be used in a CBIR system that provides a limited number of concept detectors. A query for a non-detectable concept could be solved by searching for semantically related, detectable concepts. Likewise, semantics-based systems can profit from incorporating image analysis techniques. This enables, for example, browsing facilities based on both visual and conceptual similarity between visual resources. We foresee that future research into the combination of content-based and ontology-based techniques will further enhance the process of creating semantic annotations for visual resources.

Appendix A

Guidelines for the Classification of Image Descriptions

Input texts

Text 1 "Not a day passed by without the squirrel taking a walk. He would drop himself from the beach tree on to the moss, or sometimes from the tip of a branch into the pond on the back of the dragonfly, that would take him in silence to the other side." (Translated from Dutch) [T. Tellegen. *Er ging geen dag voorbij*. Querido, Amsterdam, 1984]

Text 2 "Evening after evening the pink and yellow in the air would melt together with the green of the fields, in houses the lights were turned on and eventually it grew silent everywhere, even if it was only for a moment, because then the birds started again as the first light broke the sky."(Translated from Dutch) [G. Mak. *Het ontsnapte land*. Atlas, Amsterdam, 2001]

Text 3 "Jorritsma, Kok and VVD state secretary Van Hoof of the department of Defence spoke Saturday with the US ambassador Sobel. The Cabinet hopes to hear, this coming week, whether the Americans will give The Netherlands some more time. Kok said Friday that The Netherlands will take a decision on the JSF-project after the elections, because the political power struggle in the country will be clear then." (Translated from Dutch) [Nederlands kabinet in de wacht gezet. *Volkskrant*, page 1, May 15, 2002]

Splitting descriptions into fragments

The following guidelines are established to split image descriptions into fragments suitable for classification:

1. Separate words, as opposed to sentences, are classified separately.
2. Adjectives and Adverbs are classified together with the accompanying nouns or verbs, unless
   - the adjective conveys an action of an object (A flying | dragonfly).

- the adjective or adverb indicates color, shape, or texture.
- the adjective or adverb doesn't precede or follow the noun or verb directly, as expected in a sentence (A Squirrel, | inquisitive, | casual, | is walking...)

3. Words specifying an amount (numerals, a lot, little) are not classified separately. This follows from the fact that we don't differentiate between singular and plural words. It would be inconsistent to classify amounts separately.
4. Verbs are separate elements, unless they can be replaced by a form of the verb 'to be' without changing the meaning of the description. ("The tree stands in the forest" can be replaced by "the tree is in the forest")
5. A verb and a noun are classified as one element if the combination describes a specific, commonly occurring, action (to look at ones watch, to shake hands, to give a press conference).
6. Indefinite words like someone, somewhere, anything, are not classified as separate elements.
7. Prepositions are separate elements, unless
   - the preposition can be replaced by the phrase 'and also' ("I see a forest with a pool" can be replaced by "I see a forest and also a pool").
   - the preposition is part of a noun-verb combination as described in rule 5 (lying on ones back).
   - the preposition is part of a time specification (in spring) or a commonly used place specification (in the woods, in the air, at the table).
8. Words that are linked by 'of' (possessive) are classified as one (the back of a dragonfly).

Classification of fragments

We used three-letter-codes to classify the fragments. The first letter represents the level, the second letter represents the scope and the third letter represents a characteristic. A level is assigned to each element. Nonvisual elements are not further specified, but perceptual and conceptual elements are additionally classified by a scope and a characteristic. Characteristics are mandatory at the conceptual levels but optional at the perceptual level.

A.0.1 First Letter: Levels

Nonvisual Descriptions of the carrier or medium of the image (as opposed to the *content* of the image). This can for example be information about the author (Anton Pieck) or the style (Expressionism).

Perceptual Descriptions of the visual properties of the image. Describes that what can actually be seen in the image, what a baby would perceive or what a computer could recognize. No knowledge of the world is needed to formulate or understand perceptual descriptions. Examples are red, round, square, light, dark, but also references to the technique used to make the image, such as drawing or comic.

First letter		**Second letter**		**Third letter**	
Nonvisual	n				
Perceptual	p	Image	s	Color	l
		Element	o	Shape	s
				Texture	t
				Composition	c
				Technique	y
General	g	Scene	s	*Event*	*e*
Specific	s	Object	o	*Time*	*t*
Abstract	a			*Place*	*p*
				Relation	*r*

Table A.1 Three-letter-codes to classify description fragments.

General Descriptions of concepts that everyone is familiar with. The descriptions require only general, everyday knowledge of the world. An example is “an ape eating a banana”.

Specific Descriptions of a unique object (Wim Kok, the American Ambassador, the Dutch), or descriptions that require some form of specialist knowledge (Acute Myeloblastic Leukemia). One way to determine whether a fragment is specific, is to ask the question “Would someone from a completely different culture understand the meaning of this description?”. If the answer is No, the fragment apparently needs specialist knowledge and is classified as specific. Another way to distinguish the specific level from the general level is to look at the basic level categories of Rosch. Consider everything specific that is more specific than the concepts at this basic level. As stated in section A, adjectives are classified together with nouns. If both are general, the combination is also general. The combinations “a large house”, “a very big tree”, “a hotel room”, are all general.

Abstract Subjective knowledge is required to formulate abstract descriptions. Different abstract interpretations can be formulated of one image. Examples are despair, infinite, paradise, to beam warmth.

A.0.2 Second Letter: Scene or Object

A fragment is about a *scene* if it refers to the image as a whole (a press conference, a sunset). This are often descriptions of the light, the air, the time, or places like forests, hotel rooms. A fragment is about an *object* if it refers to a single object in the image (the prime minister, a squirrel).

A.0.3 Third Letter: Characteristics

Perceptual A perceptual fragment describes one of the following characteristics: color, shape, texture, composition and technique. Color and shape are self-evident, texture was not used by participants.

Composition The fragment describes the spatial distribution of (objects in) the image. Composition fragments are almost always prepositions. (a dragonfly *above* a pool, a squirrel *on* a branch, *in* the background, behind, next to)

Technique The fragments describes the technique that is used to produce the image (drawing, television image, puppet theater, cartoon). This characteristic is usually associated with a scene, but can also occur with objects ("a squirrel, drawn with a pencil").

Conceptual A conceptual fragment can be classified as event, place time or relation.

Event The fragment describes an action (for objects) or event(for scenes). Events are usually verbs, such as walking, talking, giving a press conference, rising of the sun. Classify all verbs as events, with the exception of forms of the verb "to be" or verbs that can be replaced by the verb "to be" without changing the meaning of the description (e.g "The three stands in the forrest" can be replaced by "The three is in the forrest").

Place The fragments describes places like Brussels, forrest, around the pool, in a meadow, in a conference room.

Time The fragments describes time specifications like eight o'clock, in spring, at sunrise. In practise, time specifications are always associated with a scene.

Relation The fragment describes the relationships between objects in the image, without specifying the composition. Examples are "walking *towards* the pool", "standing *opposite* each other", "living *in* a house". Relation fragments are almost always prepositions.

Appendix B

E-Culture Annotation Template

```
<?xml version='1.0' encoding='UTF-8'?> <!DOCTYPE rdf:RDF [
    <!ENTITY owl    'http://www.w3.org/2002/07/owl#'>
    <!ENTITY rdf    'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
    <!ENTITY rdfs   'http://www.w3.org/2000/01/rdf-schema#'>
    <!ENTITY ulan   'http://www.getty.edu/vocabularies/ulan#'>
    <!ENTITY tgn    'http://www.getty.edu/vocabularies/tgn#'>
    <!ENTITY aat    'http://www.getty.edu/vocabularies/aat#'>
    <!ENTITY vp     'http://www.getty.edu/vocabularies/vp#'>
    <!ENTITY wn     'http://wordnet.princeton.edu/wn#'>
    <!ENTITY vra    'http://www.vraweb.org/vracore/vracore3#'>
    <!ENTITY ec     'http://www.multimedian.nl/projects/n9c/eculture#'> ]>

<rdf:RDF
    xml:base="http://www.multimedian.nl/projects/n9c/eculture"
    xmlns:owl="&owl;"
    xmlns:rdf="&rdf;"
    xmlns:rdfs="&rdfs;"
    xmlns:ulan="&ulan;"
    xmlns:tgn="&tgn;"
    xmlns:aat="&aat;"
    xmlns:vp="&vp;"
    xmlns:wn="&wn;"
    xmlns:vra="&vra;"
    xmlns:ec="&ec;">

<!-- Properties used for art-historic annotation -->

<rdfs:Class rdf:ID="Image">
  <rdfs:subClassOf rdf:resource="&vra;Image" />
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&vra;relation.depicts"/>
      <owl:allValuesFrom rdf:resource="#Work"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&vra;measurements.resolution"/>
```

```
      <owl:allValuesFrom rdf:resource="&rdfs;literal"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</rdfs:Class>

<rdfs:Class rdf:ID="Work">
  <rdfs:subClassOf rdf:resource="&vra;Work" />
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&vra;relation.depictedBy"/>
      <owl:allValuesFrom rdf:resource="#Image"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&vra;creator"/>
      <owl:allValuesFrom>
        <owl:Restriction>
          <owl:onProperty rdf:resource="&vp;rolePreferred"/>
          <owl:hasValue rdf:resource="&ulan;31100"/> <!-- artist-->
        </owl:Restriction>
      </owl:allValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&vra;material.medium"/>
      <owl:allValuesFrom>
        <owl:Restriction>
          <owl:onProperty rdf:resource="&vp;parentPreferred"/>
          <owl:hasValue rdf:resource="&aat;300010358"/><!-- materials -->
        </owl:Restriction>
      </owl:allValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&vra;material.support"/>
      <owl:allValuesFrom>
        <owl:Restriction>
          <owl:onProperty rdf:resource="&vp;parentPreferred"/>
          <owl:hasValue rdf:resource="&aat;300010357"/> <!-- materials -->
        </owl:Restriction>
      </owl:allValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&vra;stylePeriod"/>
```

```
      <owl:allValuesFrom>
        <owl:Restriction>
          <owl:onProperty rdf:resource="&vp;parentPreferred"/>
          <owl:hasValue rdf:resource="&aat;300015646"/><!-- Styles and Periods -->
        </owl:Restriction>
      </owl:allValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&vra;culture"/>
      <owl:allValuesFrom rdf:resource="&ulan;Nationality"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&vra;location.creationSite"/>
      <owl:allValuesFrom rdf:resource="&tgn;Subject"/>
      <!-- all places in TGN are instances of tgn:subject. -->
    </owl:Restriction>
  </rdfs:subClassOf>
</rdfs:Class>

<!-- Classes and Properties for content annotations -->

<rdf:Property rdf:ID="subjectType" rdfs:label="subject type">
  <rdfs:domain rdf:resource="#Image"/>
  <rdfs:range>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&vp;parentPreferred"/>
      <owl:hasValue rdf:resource="&aat;300191098"/>
      <!-- visual works by subject type-->
    </owl:Restriction>
  </rdfs:range>
  <rdfs:subPropertyOf rdf:resource="&vra;subject"/>
</rdf:Property>

<rdf:Property rdf:ID="event" rdfs:label="event">
  <rdfs:domain rdf:resource="#Image"/>
  <rdfs:range>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <rdfs:Class rdf:resource="&wn;NounSynset"/>
        <rdfs:Class rdf:resource="&rdfs;literal"/>
      </owl:unionOf>
    </owl:Class>
  </rdfs:range>
  <rdfs:subPropertyOf rdf:resource="&vra;subject"/>
</rdf:Property>
```

```
<rdf:Property rdf:ID="place" rdfs:label="place">
  <rdfs:domain rdf:resource="#Image"/>
  <rdfs:range>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Restriction>
          <owl:onProperty rdf:resource="&wn;hyponymOf"/>
          <owl:hasValue rdf:resource="&wn;100022625-location-n"/>
        </owl:Restriction>
        <rdfs:Class rdf:about="&tgn;Subject"/>
        <rdfs:Class rdf:resource="&rdfs;literal"/>
      </owl:unionOf>
    </owl:Class>
  </rdfs:range>
  <rdfs:subPropertyOf rdf:resource="&vra;subject"/>
</rdf:Property>

<rdf:Property rdf:ID="time" rdfs:label="time">
  <rdfs:domain rdf:resource="#Image"/>
  <rdfs:range>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Restriction>
          <owl:onProperty rdf:resource="&wn;hyponymOf"/>
          <owl:hasValue rdf:resource="&wn;114257468-time_period-n"/>
        </owl:Restriction>
        <rdfs:Class rdf:resource="&rdfs;literal"/>
      </owl:unionOf>
    </owl:Class>
  </rdfs:range>
  <rdfs:subPropertyOf rdf:resource="&vra;subject"/>
</rdf:Property>

<rdfs:Class rdf:ID="Region"> </rdfs:Class>

<rdf:Property rdf:ID="regionDepicts">
  <rdfs:domain rdf:resource="#Region"/>
  <rdfs:range>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <rdfs:Class rdf:about="&wn;NounSynset"/>
        <rdfs:Class rdf:about="&ulan;Person"/>
        <rdfs:Class rdf:resource="&rdfs;literal"/>
      </owl:unionOf>
    </owl:Class>
  </rdfs:range>
  <rdfs:seeAlso rdf:resource="http://xmlns.com/foaf/0.1/depicts"/>
```

```
  <rdfs:subPropertyOf rdf:resource="&ec;depicts"/>
</rdf:Property>

<rdf:Property rdf:ID="regionImage">
  <rdfs:domain rdf:resource="#Region"/>
  <rdfs:range rdf:resource="#Image"/>
</rdf:Property>

<rdf:Property rdf:ID="objectDescription" rdfs:label="object description">
  <rdfs:domain rdf:resource="&vra;VisualResource"/>
  <rdfs:range rdf:resource="&ec;ObjectDescription"/>
  <rdfs:subPropertyOf rdf:resource="&ec;depicts"/>
</rdf:Property>

<rdfs:Class rdf:ID="ObjectDescription" rdfs:label="Object description"/>

<rdf:Property rdf:ID="subject" rdfs:label="subject">
  <rdfs:domain rdf:resource="#ObjectDescription"/>
  <rdfs:range rdf:resource="#Region"/>
  <rdfs:subPropertyOf rdf:resource="&ec;depicts"/>
</rdf:Property>

<rdf:Property rdf:ID="relation" rdfs:label="relation">
  <rdfs:domain rdf:resource="#ObjectDescription"/>
  <rdfs:range>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Restriction>
          <owl:onProperty rdf:resource="&wn;hyponymOf"/>
          <owl:hasValue rdf:resource="&wn;100027929-relation-n"/>
        </owl:Restriction>
        <rdfs:Class rdf:about="&wn;VerbSynset"/>
      </owl:unionOf>
    </owl:Class>
  </rdfs:range>
  <rdfs:subPropertyOf rdf:resource="&ec;depicts"/>
</rdf:Property>

<rdf:Property rdf:ID="object" rdfs:label="object">
  <rdfs:domain rdf:resource="#ObjectDescription"/>
  <rdfs:range rdf:resource="#Region"/>
  <rdfs:subPropertyOf rdf:resource="&ec;depicts"/>
</rdf:Property>

<owl:TransitiveProperty rdf:ID="depicts" rdfs:label="depicts">
  <rdfs:subPropertyOf rdf:resource="&vra;subject"/>
</owl:TransitiveProperty>
</rdf:RDF>
```

Bibliography

Abella, A. and Kender, J. (1999). From images to sentences via spatial relations. In *Proceedings of the ICCV'99 Workshop on Integration of Image and Speech Understanding*, pages 117 – 146.

Adams, B. (2003). Where does computational media aesthetics fit? *IEEE Multimedia*, **10**(2), 18–27.

Ahern, S., King, S., and Davis, M. (2005). MMM2: mobile media metadata for photo sharing. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 790 – 791.

Amin, A., Assem, van, M. Boer, de, V., Hardamn, L., HildeBrand, M., Hollink, L., Kersen, van, J., Omelayenko, B., Ossenbruggen, van, J., Schreiber, A. B., Siebes, R., Taekema, J., Wielemaker, J., and Wielinga, B. (2006). Multimedian e-culture demonstrator: objectives and architecture. Technical report, CWI/VU/UVA/DEN.

Amir, A., Berg, M., Chang, S.-F., Hsu, W., Iyengar, G., Lin, C.-Y., Naphade, M., Natsev, A., Neti, C., Nock, H., Smith, J., Tseng, B., Wu, Y., and Zhang, D. (2003). IBM research TRECVID-2003 video retrieval system. In *TREC Video Retrieval Evaluation Online Proceedings*.

Annotation Working Group (1995). Collaboration, knowledge representation and automatability. Technical report, W3C. Electronic document. Last updated May 2004. Accessed April 2006. Available from: http://www.w3.org/Collaboration/.

Antoniou, G. and Harmelen, van, F. (2004). *A semantic web primer*. MIT Press, Cambridge, MA, USA.

Armitage, L. H. and Enser, P. G. B. (1997). Analysis of user needs in image archives. *Journal of Information Science*, **23**(4), 287–299.

Assem, van, M., Menken, M. R., Schreiber, A. Th., Wielemaker, J., and Wielinga, B. J. (2004). A method for converting thesauri to RDF/OWL. In *Proceedings of the Third International Semantic Web Conference*, pages 17–31.

Assem, van, M., Gangemi, A., and Schreiber, A. B. (2006). RDF/OWL representation of WordNet. W3C editor's draft. Electronic document. Accessed May 2006. Available from: http://www.w3.org/2001/sw/BestPractices/WNET/wn-conversion.html.

Berg, van den, J. (1995). Subject retrieval in pictorial information systems. In *Proceedings of the 18th International Congress of Historical Sciences, Round Table 34: Electronic Filing, Recording, and Communication of Visual Historical Data*, pages 21–29.

Bertini, M., Del Bimbo, A., and Torniai, C. (2005). Automatic video annotation using ontologies extended visual information. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 395 – 398. Short paper.

Bloehdorn, S., Petridis, K., Saathoff, C., Simou, N., Tzouvaras, V., Avrithis, Y., Handschuh, S., Kompatsiaris, Y., Staab, S., and Strintzis, M. G. (2005). Semantic annotation of images and videos for multimedia analysis. In *Proceedings of the Second European Semantic Web Conference*, pages 592–607.

Boer, de, V., Someren, van, M., and Wielinga, B. J. (2006). Extracting instances of relations from web documents using redundancy. In *Proceedings of the Third European Semantic Web Conference*. To appear.

Boley, H., Tabet, S., and Wagner, G. (2001). Design rationale of ruleml: A markup language for semantic web rules. In *Semantic Web Working Symposium*.

Booch, G., Rumbaugh, J., and Jacobson, I. (1998). *The unified modeling language user guide*. Addison-Wesley, Reading, MA, USA.

Brickley, D. and Guha, R. V. (2000). Resource description framework (RDF) schema specification 1.0. W3C candidate recommendation. Electronic document. Accessed May 2006. Available from: http://www.w3.org/TR/rdf-schema/.

Brickley, D. and Guha, R. V. (2004). RDF vocabulary description language 1.0: RDF schema. W3C recommendation. Electronic document. Accessed May 2006. Available from: http://www.w3.org/TR/rdf-schema/.

Brink, van den, W. P. and Koele, P. (2002). *Statistiek*, volume 3. Boom, Amsterdam, The Netherlands.

Broekstra, J. and Kampman, A. (2003). SeRQL: A second generation RDF query language. In *Proceedings of the SWAD-Europe Workshop on Semantic Web Storage and Retrieval*, pages 13–14, Amsterdam, The Netherlands.

Browne, P., Czirjek, C., Gaughan, G., Gurrin, C., Jones, G. J. F., Lee, H., Marlow, S., McDonald, K., Murphy, N., OConnor, N. E., OHare, N., Smeaton, A. F., and Ye, J. (2003). Dublin City University video track experiments for TREC 2003. In *TREC Video Retrieval Evaluation Online Proceedings*.

Buchanan, B. G. and Shortliffe, E. H., editors (1984). *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA, USA.

Carletta, J. (1996). Asessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, **22**(2), 249–254.

Chang, S. F., Chen, W., and Sundaram, H. (1998). Semantic Visual Templates: linking visual features to semantics. In *IEEE International Conference on Image Processing (ICIP '98)*, pages 531–535, Chicago, Illinois.

Chen, H. (2001). An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology*, **52**(3), 260–273.

Choi, Y. and Rasmussen, E. M. (2002). Users' relevance criteria in image retrieval in american

history. *Information Processing and Management*, **38**(5), 695–726.

Christel, M. and Moraveji, N. (2004). Finding the right shots: assessing usability and performance of a digital video library interface. In *Proceedings of ACM Multimedia*, pages 732–739.

Clayphan, R. and Oldroyd, B. (2005). Using Dublin Core application profiles to manage diverse metadata developments. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*.

Cohn, A. G. and Hazarika, S. M. (2001). Qualitative spatial representation and reasoning: an overview. *Fundamenta Informaticae*, **46**(1-2), 1–29.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**(6), 391–407.

Doerr, M. (2003). The CIDOC CRM - An ontological approach to semantic interoperability of metadata. *AI Magazine*, **24**(3), 75–92.

Dublin Core (2006). Dublin Core metadata element set, version 1.1: reference description. Dublin Core Metadata Initiative. Electronic document. Accessed January 2006. Available from: http://dublincore.org/documents/dces/.

E-Culture (2006). Multimedian n9c eculture project homepage. Electronic document. Accessed May 2006. Available from: http://e-culture.multimedian.nl/.

Eakins, J. P. (1998). Techniques for image retrieval. *Library and Information Briefings*, **85**, 1–15.

Eakins, J. P. (2002). Towards intelligent image retrieval. *Pattern Recognition*, **35**(1), 3–14.

Eakins, J. P., Briggs, P., and Burford, B. (2004). Image retrieval interfaces: a user perspective. In *Proceedings of the Third International Conference on Image and Video Retrieval*, pages 628–637.

Enser, P. G. B. and McGregor, C. G. (1992). Analysis of visual information retrieval queries. Research and Development Report 6104, British Library.

Fang, X. and Salvendy, G. (2000). Keyword comparison: a user-centered feature for improving web search tools. *International Journal of Human-Computer Studies*, **52**(5), 915–931.

Fellbaum, C., editor (1998). *WordNet: an electronic lexical database*, Cambridge, MA, USA. MIT press.

Fidel, R. (1997). The image retrieval task: implications for the design and evaluation of image databases. *The New Review of Hypermedia and Multimedia*, **3**, 181–199.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. (1995). Query by image and video content: the qbic system. *IEEE Computer*, **28**(9), 23–32.

Frost, C. O. and Noakes, A. (1998). Browsing images using broad classification categories. In *Proceedings of te ninth ASIS SIGCR Classification Research Workshop*, pages 71–89.

Gauvain, J., Lamel, L., and Adda, G. (2002). The limsi broadcast news transcription system.

Speech Communication, **37**(1-2), 89–108.

Getty Foundation, The (2006a). The art and architecture thesaurus (AAT). Electronic document. Accessed March 2006. Available from: http://www.getty.edu/research/conducting_research/vocabularies/aat/.

Getty Foundation, The (2006b). The getty vocabulary program. Electronic document. Accessed March 2006. Available from: http://www.getty.edu/research/conducting_research/vocabularies/.

Getty Foundation, The (2006c). The thesaurus of geographical names (TGN). Electronic document. Accessed March 2006. Available from: http://www.getty.edu/research/conducting_research/vocabularies/tgn/.

Getty Foundation, The (2006d). The union list of artist names (ULAN). Electronic document. Accessed March 2006. Available from: http://www.getty.edu/research/conducting_research/vocabularies/ulan/.

Geusebroek, J. M. and Smeulders, A. W. M. (2005). A six-stimulus theory for stochastic texture. *International Journal of Computer Vision*, **62**(1/2), 7–16.

Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP*, pages 38 – 44.

Goodrum, A. and Spink, A. (2001). Image searching on the excite web search engine. *Information Processing and Management*, **37**(2), 295–311.

Goodrum, A., Bejune, M. M., and Siochi, A. C. (2003). A state transition analysis of image search patterns on the web. In *Proceedings of the Second International Conference on Image and Video Retrieval*, volume 2728, pages 281–290.

Graham, M. E. (1999). The description and indexing of images: report of a survey of arlis members. Technical report, Institute for Image Data Research, University of Northumbria at Newcastle. Electronic document. Accessed January 2002.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**(2), 199–220.

Gudivada, V. N. and Raghavan, V. V. (1995). Content-based image retrieval systems. *IEEE Computer*, **28**(9), 18–22.

Gupta, A. and Jain, R. (1997). Visual information retrieval. *Communications of the ACM*, **40**(5), 70–79.

Halaschek-Wiener, C., Schain, A., Golbeck, J., Grove, M., Parsia, B., and Hendler, J. (2005). A flexible approach for managing digital images on the semantic web. In *Proceedings of the Fifth International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot)*, pages 49–58.

Handschuh, S. and Staab, S., editors (2003a). *Annotation for the semantic web*, volume 96 of *Frontiers in Artificial Intelligence and Applications*, Amsterdam, The Netherlands. IOS Press.

Handschuh, S. and Staab, S. (2003b). Annotation of the shallow and the deep web. In S. Handschuh and S. Staab, editors, *Annotation for the semantic web*, volume 96 of *Frontiers in Artificial Intelligence and Applications*, page 25. IOS Press.

Harden, M. (2006). Mark harden's artchive. Electronic document. Accessed April 2006. Available from: http://www.artchive.com/.

Hatala, M. and Richards, G. (2003). Value-added Metatagging: Ontology and Rule based Methods for Smarter Metadata. In *Rules and Rule Markup Languages for the Semantic Web (RuleML2003)*, pages 65–80.

Hauptmann, A., Baron, R. V., Chen, M.-Y., Christel, M., Duygulu, P., Huang, C., Jin, R., Lin, W.-H., Ng, T., Moraveji, N., Papernick, N., Snoek, C. G. M., Tzanetakis, G., Yang, J., Yan, R., and Wactlar, H. D. (2003). Informedia at TRECVID 2003: analyzing and searching broadcast news video. In *TREC Video Retrieval Evaluation Online Proceedings*.

Hauptmann, A. G. (2004). Towards a large scale concept ontology for broadcast video. In *Proceedings of the Third International Conference on Image and Video Retrieval*, pages 674–675.

Hauptmann, A. G. and Christel, M. G. (2004). Successful approaches in the TREC video retrieval evaluations. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 668–675.

Hecht, A., O'Dwyer, A., Oomen, J., and Scharinger, F. (2004). Birth: building an interactive research and delivery network for television heritage. In *Proceedings of the International Cultural Heritage Informatics Meeting*.

Heesch, D., Pickering, M. J., Ruger, S., and Yavlinsky, A. (2003). Video retrieval using search and browsing with keyframes. In *TREC Video Retrieval Evaluation Online Proceedings*.

Heidorn, P. B. (1999). The identification of index terms in natural language object descriptions. In *Proceedings of the American Society for Information Science Annual Meeting*, volume 36, pages 472–481.

Hillmann, D. (2001). Using dublin core. Recomendation, Dublin Core Metadata Initiative.

Hoang, M. A., Geusebroek, J., and Smeulders, A. W. M. (2002). Color texture measurement and segmentation. In *Proceedings of the second international workshop on Texture Analysis and Synthesis*, pages 73–76.

Hollink, L., Schreiber, A. Th., Wielemaker, J., and Wielinga, B. J. (2003). Semantic annotation of image collections. In *Proceedings of the K-Cap 2003 Workshop on Knowledge Markup and Semantic Annotation*.

Hollink, L., Nguyen, G., Schreiber, A. B., Wielemaker, J., Wielinga, B. J., and Worring, M. (2004a). Adding spatial semantics to image annotations. In *Proceedings of the fourth International Workshop on Knowledge Markup and Semantic Annotation at ISWC*, pages 31–40.

Hollink, L., Schreiber, A. Th., Wielinga, B. J., and Worring, M. (2004b). Classification of user image descriptions. *International Journal of Human Computer Studies*, **61**(5), 601–621.

Hollink, L., Nguyen, G. P., Koelma, D. C., Schreiber, A. Th., and Worring, M. (2004c). User strategies in video retrieval: a case study. In *Proceedings of the Third International Conference on Image and Video Processing*, pages 6 – 14, Dublin, Ireland.

Hollink, L., Nguyen, G. P., Koelma, D., Schreiber, A. B., and Worring, M. (2005a). Assessing user behaviour in news video retrieval. *IEE proceedings on Vision, Image and Signal Processing*, **152**(6), 911–918.

Hollink, L., Worring, M., and Schreiber, A. B. (2005b). Building a visual ontology for video retrieval. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 479 – 482. Short paper.

Hollink, L., Little, S., and Hunter, J. (2005c). Evaluating the application of semantic inferencing rules to image annotation. In *Proceedings of the Third International Conference on Knowledge Capture*, pages 91 – 98, Banff, Canada.

Hoogs, A., Rittscher, J., Stein, G., and Schmiederer, J. (2003). Video content annotation using visual analysis and a large semantic knowledge base. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 327 – 334.

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., and Dean, M. (2004). SWRL: A semantic web rule language combining OWL and RuleML. W3C member submission. Electronic document. Accessed May 2006. Available from: http://www.w3.org/Submission/SWRL/.

Houten, van, Y., Schuurman, J. G., and Verhagen, P. (2004). Video content foraging. In *Proceedings of the Third International Conference on Image and Video Retrieval*, pages 15–23.

Hunter, J. (2001). Adding multimedia to the semantic web - building an MPEG-7 ontology. In *International Semantic Web Working Symposium (SWWS)*.

Hunter, J., Drennan, J., and Little, S. (2004). Realizing the Hydrogen Economy through Semantic Web Technologies. *IEEE Intelligent Systems Journal - Special Issue on eScience*, **19**(1), 40–47.

Hyvönen, E., Kettula, S., Raatikka, V., Saarela, S., and Viljanen, K. (2003). Finnish museums on the semantic web. In *Proceedings of the Twelfth International World Wide Web Conference*, Budapest, Hungary. Poster.

Hyvönen, E., Salminen, M., Junnila, M., and Kettula, S. (2004a). A content creation process for the semantic web. In *Proceedings of the LREC Workshop on Ontologies and Lexical Resources in Distributed Environments*.

Hyvönen, E., Saarela, S., Viljanen, K., Mäkelä, E., Valo, A., Salminen, M., Kettula, S., and Junnila, M. (2004b). A cultural community portal for publishing museum collections on the semantic web. In *Proceedings of the ECAI Workshop on Application of Semantic Web Technologies to Web*.

Hyvönen, E., Junnila, M., Kettula, S., Mäkelä, E., Saarela, S., Salminen, M., Syreeni, A., Valo, A., , and Viljanen, K. (2004c). Finnish museums on the semantic web: the users perspective on MuseumFinland. In *Proceedings of Museums on the Web*, pages 21–32.

IFLA (1998). Functional requirements for bibliographic records. Final report, International Federation of Library Associations and Institutions, München, Germany. UBCIM publication - New Series, volume 19.

Jaimes, A. and Chang, S.-F. (2000). A conceptual framework for indexing visual information at multiple levels. In *Proceedings of SPIE Internet Imaging*, volume 3964.

Jaimes, A., Tseng, B. L., and Smith, J. R. (2003). Modal keywords, ontologies, and reasoning for video understanding. In E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. Zhou, editors, *Proceedings of the International Conference on Image and Video Retrieval*, pages 248–259, Urbana, IL, USA.

Jansen, B. J., Goodrum, A., and Spink, A. (2000). Searching for multimedia: analysis of audio, video and image web queries. *World Wide Web*, **3**(4), 249–254.

Jörgensen, C. (1996). Indexing images: testing an image description template. In *Proceeding of The American Society for Information Science Annual Conference*.

Jörgensen, C. (1998). Attributes of images in describing tasks. *Information Processing and Management*, **34**(2/3), 161–174.

Jörgensen, C. (1999). Image indexing - an analysis of selected classification systems in relation to image attributes named by naive users. *Annual Review of OCLC Research*.

Kettler, B., Starz, J., Miller, W., and Haglich, P. (2005). A template-based markup tool for semantic web content. In Y. Gill, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *Proceedings of the 4th International Semantic Web Conference*, pages 446–460.

Koivunen, M.-R. (2005). Annotea project. Electronic document. Accessed April 2006. Available from http://www.w3.org/2001/Annotea/.

Koivunen, M.-R. and Swick, R. R. (2003). Collaboration through annotation on the semantic web. In S. Handschuh and S. Staab, editors, *Annotation for the Semantic Web*, volume 96 of *Frontiers in Artificial Intelligence and Applications*, page 4660. IOS Press.

Lafon, Y. and Bos, B. (2002). Describing and retrieving photographs using RDF and HTTP. W3C note. Electronic document. Accessed April 2006. Available from: http://www.w3.org/TR/photo-rdf/.

Lahti, J., Westermann, U., Palola, M., Petlola, J., and Vildjiounaite, E. (2005). Integrated capture, annotation, and sharing of video clips with mobile phones. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 790 – 791.

Lee, E. (2004). Building interoperability for united kingdom environment information resources. In *Proceedings of the 8th European Conference on Digital Libraries*, pages 179–185.

Lee, S. and Hwang, E. (2002). Spatial similarity and annotation-based image retrieval system. In *Proceedings of the IEEE Fourth International Symposium on Multimedia Software Engineering*, pages 33–37.

Ley, J. (2004). Raster image description and search in SVG. Presented at the third annual confer-

ence on Scalable Vector Graphics (SVG Open). Electronic document. Accessed March 2005. Available from http://www.jibbering.com/svg/talk2004/title.html.

Little, S. and Hunter, J. (2004). Rules-By-Example - a novel approach to semantic indexing and querying of images. In *Proceedings of the Third International Semantic Web Conference*, pages 534–548.

Liu, S., Liu, F., Yu, C., and Meng, W. (2004). An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 266–272.

Marques, O. and Barman, N. (2003). Semi-automatic Semantic Annotation of Images Using Machine Learning Techniques. In D. Fensel, K. Sycara, and J. Mylopoulos, editors, *Proceedings of the International Semantic Web Conference*, pages 550–565, Florida.

Marsh, B. J., Mastronarde, D. N., Buttle, K. F., Howell, K. E., and McIntosh, J. R. (2001). Organellar relationships in the golgi region of the pancreatic beta cell line, HIT-T15, visualized by high resolution electron tomography. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(5), 2399–2406.

Martínez, J. M. (2001). Overview of the MPEG-7 standard. Technical Report 5.0, ISO/IEC.

Martínez, J. M., González, C., Fernández, O., García, C., and J., R. (2002). Towards universal access to content using MPEG-7. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 199–202.

Mezaris, V., Kompatsiaris, I., and Strintzis, M. G. (2004). Region-based image retrieval using an object ontology and relevance feedback. *European Association for Signal, Speech and Image Processing Journal on Applied Signal Processing*, **6**, 886 901.

MIA (2002). Homepage of the Multimedian Information Analysis project. Electronic document. Accessed May 2006. Available form: http://www.ins.cwi.nl/projects/MIA/.

Moldovan, D. I. and Mihalcea, R. (2000). Using WordNet and lexical operators to improve internet searches. *IEEE Internet Computing*, **4**(1), 34–43.

Naphade, M. and Huang, T. (2001). Detecting semantic concepts using context and audiovisual features. In *Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video, 2001*, pages 92–98.

Navigli, R. and Velardi, P. (2003). An analysis of ontology-based query expansion strategies. In *Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining*, pages 42–49, Cavtat-Dubrovnik, Croatia.

Nguyen, G. P. and Worring, M. (2003). Query definition using interactive saliency. In *Proceedings of the 5th ACM Multimedia International Workshop on Multimedia Information Retrieval*, pages 150–156, Berkeley, CA, USA.

Nguyen, G. P. and Worring, M. (2004). Optimizing similarity based visualization in content based image retrieval. In *Proceeding of the IEEE ICME special session on Novel Techniques for*

Browsing in Large Multimedia Collections, pages 759–762, Taipei, Taiwan.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In C. Welty and B. Smith, editors, *Proceedings of The International Conference on Formal Ontology in Information Systems*, pages 2–9, Ogunquit, ME, USA.

Niles, I. and Pease, A. (2003). Linking lexicons and ontologies: mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering*, pages 23–36.

NIST (2005). *Guidelines for the TRECVID 2004 evaluation*. Electronic document. Last updated February 2005. Accessed April 2006. Available from: http://www-nlpir.nist.gov/projects/tv2004/tv2004.html.

O'Hare, N., Gurrin, C., Lee, H., Murphy, N., Smeaton, A. F., and Jones, G. J. F. (2005). My digital photos: where and when? In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 790–791.

Ornager, S. (1995). The newspaper image database: empirical supported analysis of users' typology and word association clusters. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, volume 3964, pages 212–218.

Panofsky, E. (1962). *Studies in iconology*. Harper and Row, New York, NY, USA.

Rashid, A., Shariff, B. M., and Egenhofer, M. J. (1998). Natural-language spatial relations between linear and areal objects: the topology and metric of english-language terms. *International Journal of Geographic Information Science*, **12**(3), 215–246.

Rautiainen, M., Penttilä, J., Pietarila, P., Noponen, K., Hosio, M., Koskela, T., Mäkelä, S.-M., Peltola, J., Liu, J., Ojala, T., and Seppänen, T. (2003). TRECVID 2003 experiments at MediaTeam Oulu and VTT. In *TREC Video Retrieval Evaluation Online Proceedings*.

Rosch, E. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**(3), 382–439.

Schreiber, A. Th., Dubbeldam, B., Wielemaker, J., and Wielinga, B. J. (2001). Ontology based photo annotation. *IEEE Intelligent Systems*, **16**(3), 66–74.

Schreiber, A. Th., Blok, I., Carlier, D., Gent, van, W. P. C., Hokstam, J., and Roos, U. (2002). A mini-experiment in semantic annotation. In I. Horrocks and J. Hendler, editors, *Proceedings of the First International Semantic Web Conference*, page 404408, Sardinia, Italy.

Shatford, S. (1986). Analyzing the subject of a picture: a theoretical approach. *Cataloging and Classification Quarterly*, **6**(3), 39–62.

Sim, K. M. (2004). Toward an ontology-enhanced information filtering agent. *ACM SIGMOD Record*, **33**(1), 95–100.

Sinclair, P., Lewis, P., Martinez, K., Addis, M., Prideaux, D., Fina, D., and Bormida, G. D. (2005). eCHASE: sustainable exploitation of electronic cultural heritage. In *Proceedings of the Second European Workshop on the Integration of Knowledge, Semantic and Digital Media Technolo-*

gies.

Smeaton, A. F. and Quigley, I. (1996). Experiments on using semantic distances between words in image caption retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 174–180.

Smeaton, A. F., Kraaij, W., and Over, P. (2003). TRECVID - an overview. In *TREC Video Retrieval Evaluation Online Proceedings*.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(12), 1349–1380.

Smith, B. and Welty, C. (2001). Ontology: Towards a new synthesis. In C. Welty and B. Smith, editors, *Formal Ontology in Information Systems*, pages iii–x.

Smith, J. R. and Chang, S.-F. (1996). VisualSEEk: a fully automated content-based image query system. In *Proceedings of Fourth ACM International Conference on Multimedia*, pages 87–98. ACM Press.

Smith, J. R., Campbell, M., Naphade, M., Natsev, A., and Tesic, J. (2005). Learning and classification of semantic concepts in broadcast video. In *Online Proceedings of the First International Conference on Intelligence Analysis*, McLean, VA, USA. Electronic document. Accessed May 2006. Available from: https://analysis.mitre.org/proceedings/Final_Papers_Files/362_Camera_Ready_Paper.pdf.

Snoek, C. G., Worring, M., Geusebroek, J.-M., Koelma, D. C., Seinstra, F. J., and Smeulders, A. W. (2006). The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **In Press**.

Snoek, C. G. M., Worring, M., Geusebroek, J. M., Koelma, D. C., and Seinstra, F. J. (2004). The MediaMill TRECVID 2004 semantic video search engine. In *TREC Video Retrieval Evaluation Online Proceedings*.

Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks/Cole.

Stein, G. C., Rittscher, J., and Hoogs, A. (2003). Enabling video annotation using a semantic database extended with visual knowledge. In *Proceedings of The IEEE International Conference on Multimedia and Expo*.

Talmy, L. (1983). How language structures space. In H. Pick and L. Acredols, editors, *Spatial orientation: theory, research and application*, New York, NY, USA. Plenum Press.

Tam, A. M. and Leung, C. H. C. (2002). Structured natural-language description for semantic content retrieval. *Journal of the American Society for Information Science*, **52**(11), 930–937.

Tansley, R., Bird, C., Hall, W., Lewis, P., and Weal, P. (2000). Automating the linking of content and concept. In *Proceedings of the 8th annual ACM international conference on Multimedia*, pages 445–447. ACM.

Tansley, R. H. (2000). *The Multimedia Thesaurus: Adding a Semantic Layer to Multimedia Information*. phd, University of Southhampton.

Vakkari, P. (2000). Cognition and changes of search terms and tactics during task performance; a longitudinal study. In *Proceedings of La Conférence Recherche d'Information Assistée par Ordinateur*, pages 894–907.

Volkmer, T., Smith, R., and Natsev, A. (2005). A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. In H. Zhang and T.-S. Chua, editors, *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 892–901, New York, NY, USA. ACM Press.

Voorhees, E. (1994). Query expansion using lexical-semantic relations. In W. B. Croft and C. J. Rijsbergen, van, editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, New York, NY, USA. Springer-Verlag.

VRA (2002). VRA core categories, version 3.0. Technical report, Visual Resources Association. Electronic document. Accessed April 2006. Available from: http://www.vraweb.org/vracore3.htm.

Web Ontology Working Group (2003). Ontology language overview. W3C candidate recommendation. Electronic document. Accessed April 2006. Available from: http://www.w3.org/TR/owl-features/.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Wielemaker, J. (2000). Swi-prolog rdf parser. Technical report, SWI, University of Amsterdam. Electronic document. Accessed March 2006. Available from: http://www.swi-prolog.org/packages/rdf2pl.html.

Wielemaker, J. (2005). An optimised semantic web query language implementation in prolog. In *ICLP 2005*, pages 128–142.

Wielemaker, J., Schreiber, A. Th., and Wielinga, B. J. (2003). Prolog-based infrastructure for RDF: performance and scalability. In *Proceedings of The 2nd International Semantic Web Conference*.

Wielinga, B. J., Schreiber, A. Th., Wielemaker, J., and Sandberg, J. A. C. (2001). From thesaurus to ontology. In Y. Gil, M. Musen, and J. Shavlik, editors, *Proceedings of The 1st International Conference on Knowledge Capture*, pages 21–23. ACM Press.

Worring, M., Nguyen, G. P., Hollink, L., Gemert, J., and Koelma, D. C. (2004). Accessing video archives using interactive search. In *Proceedings of the International Conference on Multimedia and Expo*, pages 297–300.

Yang, M., Wildemuth, B. M., and Marchionini, G. (2004). The relative effectiveness of concept-based versus content-based video retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 368 – 371.

Zhao, R. and Grosky, W. (2002). Negotiating The Semantic Gap: From Feature Maps to Semantic Landscapes. *Pattern Recognition*, **35**(3), 51–58.

Samenvatting: Semantische Annotatie voor het Zoeken naar Beeldmateriaal

Beeldmateriaal speelt een grote rol als onderdeel van ons cutlureel erfgoed, in de wetenschap, in het onderwijs, kortom, in de maatschappij. Zowel het aanbod van, als de vraag naar beeldmateriaal groeit. Zoeken in grote collecties beeldmateriaal blijft echter een moeizaam proces. Het kost een eindgebruiker veel tijd en moeite om juist dat ene beeld te vinden in een grote collectie. Daarom zijn er efficiënte zoekmethoden nodig om de groeiende collecties doorzoekbaar te maken en te houden. Dit proefschrift gaat over de problemen die er zijn bij het zoeken naar beeldmateriaal, en mogelijke oplossingen daarvoor. We onderzoeken dit in drie uiteenlopende domeinen: schilderijen, microscoopbeelden van organische cellen en video's van nieuwsuitzendingen.

De oudste manier om collecties van beeldmateriaal te ontsluiten is door middel van annotatie. Een annotatie is informatie die expliciet aan een item in een collectie is gekoppeld, met de bedoeling het item te beschrijven zodat het in de toekomst teruggevonden kan worden. Het is niet eenvoudig te bepalen welke informatie een annotatie moet bevatten. Immers, een plaatje zegt meer dan duizend woorden. Eén beeld kan verschillend geïnterpreteerd worden door verschillende mensen. Een foto van de val van de Berlijnse muur kan bijvoorbeeld beschreven worden als "een zwart-wit foto gemaakt door A. N. de Wit in november 1989", "mensen, nacht, beton", "een man die een muur op klimt" of "het IJzeren Gordijn". Idealiter bevat een annotatie van een beeld alle zoekvragen die mensen zouden kunnen formuleren om dat beeld terug te vinden.

Om te beginnen brengen we de verschillende manieren waarop een beeld kan worden beschreven in kaart. We maken een raamwerk, gebaseerd op literatuur, waarin beschrijvingen van beelden gecategoriseerd kunnen worden. Het raamwerk onderscheidt non-visuele beschrijvingen (zoals de maker van het beeld of de datum waarop het gemaakt is), perceptuele beschrijvingen (zoals kleuren of vormen), en conceptuele beschrijvingen (zoals de objecten, scènes en locaties die zijn uitgebeeld). Conceptuele beschrijvingen zijn verder onderverdeeld in algemene, specifieke en abstracte beschrijvingen. Om te bepalen welke categorieën van het raamwerk het meest voorkomen, gebruiken we het raamwerk voor categorisatie van beschrijvingen in drie domeinen. We zien dat mensen die beelden beschrijven twee keer zoveel beschrijvingen geven van objecten als van scènes. Algemene beschrijvingen worden verreweg het meest gebruikt (74 %), gevolgd door specifieke beschrijvingen (16 %) en abstracte beschrijvingen (9 %). Veelgebruikte subcategorieën zijn gebeurtenissen, locaties en spatiële relaties tussen objecten. In het nieuwsdomein is de 'specifieke' categorie belangrijker dan in andere domeinen.

Om te zorgen voor eenduidigheid in het brede scala aan beschrijvingen van annotators en zoekers, gebruiken cultureel erfgoed instellingen veelal gecontroleerde vocabulaires. Een gecontroleerd vocabulaire kan simpelweg een woordenlijst zijn met de termen die een annotator of

zoeker mag gebruiken. Het kan ook een meer gestructureerd geheel zijn, zoals een thesaurus of een ontologie, waarin de onderlinge (hiërarchische) relaties tussen de concepten vastgelegd zijn. Het gebruik van gestandaardiseerde en veelgebruikte vocabulaires maakt de geannoteerde collectie toegankelijk voor een breder publiek dan wanneer 'eigen' vocabulaires gebruikt worden.

Inmiddels is er veel vooruitgang geboekt in het onderzoek naar het representeren van kennis, in het bijzonder in het veld van *the semantic web*, of 'het semantische web'. Dit heeft geleid tot gestandaardiseerde talen waarin concepten en relaties formeel beschreven worden. Het formele karakter van deze talen en de vastgelegde betekenis van de onderdelen ervan, maakt het voor zowel computers als mensen mogelijk informatie uit te wisselen, te verwerken en ermee te redeneren. Wanneer ontologieën in deze *semantic web* talen worden gerepresenteerd, kunnen de concepten in de ontologieën uniek geïndentificeerd worden en krijgen relaties tussen concepten expliciete betekenis. Met andere woorden, de kennis in de ontologie kan nu worden uitgewisseld, verwerkt en er kan mee worden geredeneerd door mensen en computers. Onze hypothese is dat deze ontologieën niet alleen gebruikt kunnen worden als gecontroleerde vocabulaires, maar dat de achtergrondkennis die zij in zich hebben ook gebruikt kan worden om annotatie- en zoekconcepten te disambigueren, om het annotatie proces te versnellen en om betere zoekresultaten te behalen.

Om het gebruik van ontologieën voor het creëren van semantische annotaties te demonstreren, definiëren we ten eerste een metadata schema voor het schilderijendomein dat gebaseerd is op ons categorisatie raamwerk en bestaande standaarden. Meerdere ontologieën vormen hierbij samen het gecontroleerde vocabulaire. De velden in het metadata schema zijn gekoppeld aan relevante delen van het vocabulaire, zodat het voor de annotator of zoeker gemakkelijker is om de juiste concepten te vinden. Het 'maker' veld is bijvoorbeeld gekoppeld aan een thesaurus voor namen van kunstenaars. Het gebruik van meerdere grote vocabulaires leidt tot homoniemen. We laten zien hoe de hiërarchische relaties in een ontologie de gebruiker kunnen helpen bij het disambigueren van homonieme annotatie- of zoektermen: door de plaats van de term in de hiërarchie te tonen wordt de betekenis duidelijk. De achtergrondkennis in een ontologie kan het annotatieproces versnellen door waarden van annotaties te suggereren. In een *use case* laten we zien dat met bestaande kennis van het kunstdomein en aantal annotaties afgeleid kan worden van de naam van de maker van een schilderij: de stijl, de cultuur en het materiaal. Ten slotte kunnen relaties in een ontologie gebruikt worden om een zoekvraag uit te breiden; iemand die zoekt naar kubistische schilderijen is vast geïnteresseerd in schilderijen geannoteerd met 'Picasso' en iemand die zoekt naar bloemen zal tevreden zijn met beelden van zonnebloemen. In deze twee voorbeelden zijn respectievelijk schilder-stijl en subklasse relaties gebruikt om de zoekvraag uit te breiden. Niet alle typen relaties zijn hiervoor geschikt: een zoekvraag naar Frans Hals die beantwoord wordt met plaatjes van zijn geboortestad Haarlem zal wellicht niet bevredigend zijn. In een experimentele opzet onderzoeken we welke relaties het meest geschikt zijn om de zoekvraag uit te breiden. Een combinatie van subklasse relaties en deel-geheel relaties blijkt geschikt. Ook superklasse relaties kunnen gebruikt worden, maar niet in combinatie met andere relaties. Tot vier subklasse of deel-geheel relaties achter elkaar leveren goede resultaten op.

Een geheel andere manier om collecties van beeldmateriaal te ontsluiten is door analyse van de

visuele eigenschappen van het beeld. Met deze methode worden annotaties automatisch afgeleid uit eigenschappen zoals kleur, textuur en vorm. Grote trainingcollecties worden gebruikt om combinaties van eigenschappen te verbinden aan annotatieconcepten. Het voordeel van deze methode is dat er geen tijdrovende handmatige annotatie nodig is om collecties doorzoekbaar te maken. Zoekmethoden die gebaseerd zijn op de visuele eigenschappen van de inhoud van een beeld worden *Content Based Image Retrieval (CBIR)* methoden genoemend, of methoden voor 'inhoudsgebaseerd zoeken naar beelden'.

Het belangrijkste probleem van CBIR staat bekend als de *semantic gap*, of de 'semantische kloof'. Het gaat hier om de kloof tussen de informatie die uit de visuele gegevens gehaald kan worden en de interpretatie die een gebruiker heeft van een beeld. Abstracte beschrijvingen zoals 'het IJzeren Gordijn' kunnen bijvoorbeeld niet afgeleid worden uit de visuele eigenschappen alleen. Ook meer algemene beschrijvingen zoals 'een man die over een muur klimt' kunnen met de huidige CBIR technieken niet afgeleid worden. We onderzoeken onder welke omstandigheden de semantische kloof het zoeken het meest belemmert. Dit kan ons later helpen bij het bepalen waar CBIR het beste ingezet kan worden. We zien dat de kloof alleen overbrugd kan worden als de beelden goed gesegmenteerd kunnen worden in objecten (*strong segmentation*), als het domein klein en goed begrepen is, en als verschillende typen objecten visueel sterk van elkaar verschillen. Schilderijen noch nieuwsuitzendingen voldoen aan deze criteria. Toch zijn er in deze domeinen belangrijke resultaten te behalen met CBIR. We laten zien dat bijvoorbeeld spatiële relaties tussen objecten op een schilderij goed af te leiden zijn uit de visuele eigenschappen van het beeld. Voor nieuwsuitzendingen geldt dat over het algemeen betere zoekresultaten worden behaald door te zoeken in de tekst (o.a. de spraak en de ondertiteling), dan door CBIR te gebruiken. Als we echter de zoekvragen opdelen in algemene vragen en specifieke vragen zien we dat de eerste categorie veel baat heeft bij CBIR. Een verklaring hiervoor is dat alledaagse onderwerpen vaak te zien zijn in het nieuws, maar zelden genoemd worden door de nieuwslezer of door andere sprekers. Dit gegeven kan gebruikt worden door zoekmachines: een (automatische of handmatige) categorisatie van de zoekvraag kan gebruikt worden om de beste zoekmethode te kiezen.

Onze hypothese is dat een combinatie van CBIR en ontologie-gebaseerde methoden kwalitatief goede zoekmogelijkheden oplevert en tegelijkertijd de benodigde annotatietijd beperkt. We maken deze combinatie in de drie domeinen, op drie verschillende manieren. In het kleine, door experts goed begrepen domein van organische cellen vragen we domein experts (biologen) om regels te formuleren die domeinconcepten uit een bestaande ontologie koppelen aan visuele eigenschappen. De regels modeleren juist dat deel van de domeinkennis dat niet aanwezig is in de ontologie, namelijk de visuele eigenschappen van de concepten. Deze kennis kan vervolgens gebruikt worden bij het zoeken. Voor een subset van het nieuwsdomein maken we een 'visuele ontologie': een ontologie die zowel algemene concepten (fiets, schip) als visuele eigenschappen van deze concepten (kleuren, materialen) bevat. Dit kan bijvoorbeeld gebruikt worden om de lijst van mogelijke annotatieconcepten te beperken tot de concepten waarvan de visuele eigenschappen overeenkomen met de visuele eigenschappen van het beeld. In het domein van schilderijen kan het uiterlijk van objecten sterk wisselen. Het is hier dan ook niet haalbaar om visuele eigenschappen

direct te koppelen aan concepten in een ontologie. In plaats daarvan gebruiken we ontologieën en CBIR naast elkaar, zodat ze elkaar aanvullen. Delen van de annotatie die geschikt zijn voor CBIR technieken worden automatisch ingevuld, terwijl delen van de annotatie waarvoor een hoger niveau nodig is door een menselijke annotator ingevuld worden, waar mogelijk met behulp van een ontologie.

In dit proefschrift hebben we laten zien dat een combinatie van ontologie-gebaseerde methoden en CBIR de mogelijkheid biedt om semantische annotaties van beelden op een efficiente manier te produceren. Een voorbeeld van toekomstig onderzoek in deze richting is een CBIR systeem dat een beperkt aantal concepten automatisch kan detecteren in een beeld. Een zoekvraag naar een niet-detecteerbaar concept kan beantwoord worden door te zoeken naar gerelateerde detecteerbare concepten, waarbij de relaties tussen concepten uit een ontologie worden gehaald. Een ander voorbeeld is een browser waarin gebruikers kunnen browsen naar zowel visueel als conceptueel gerelateerde beelden.

SIKS Dissertation Series

1998-1 Johan van den Akker (CWI)
DEGAS - An Active, Temporal Database of Autonomous Objects

1998-2 Floris Wiesman (UM)
Information Retrieval by Graphically Browsing Meta-Information

1998-3 Ans Steuten (TUD)
A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective

1998-4 Dennis Breuker (UM)
Memory versus Search in Games

1998-5 E.W.Oskamp (RUL)
Computerondersteuning bij Straftoemeting

1999-1 Mark Sloof (VU)
Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products

1999-2 Rob Potharst (EUR)
Classification using decision trees and neural nets

1999-3 Don Beal (UM)
The Nature of Minimax Search

1999-4 Jacques Penders (UM)
The practical Art of Moving Physical Objects

1999-5 Aldo de Moor (KUB)
Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems

1999-6 Niek J.E. Wijngaards (VU)
Re-design of compositional systems

1999-7 David Spelt (UT)
Verification support for object database design

1999-8 Jacques H.J. Lenting (UM)
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation

2000-1 Frank Niessink (VU)
Perspectives on Improving Software Maintenance

2000-2 Koen Holtman (TUE)
Prototyping of CMS Storage Management

2000-3 Carolien M.T. Metselaar (UVA)
Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief

2000-4 Geert de Haan (VU)
ETAG, A Formal Model of Competence Knowledge for User Interface Design

2000-5 Ruud van der Pol (UM)
Knowledge-based Query Formulation in Information Retrieval

2000-6 Rogier van Eijk (UU)
Programming Languages for Agent Communication

2000-7 Niels Peek (UU)
Decision-theoretic Planning of Clinical Patient Management

2000-8 Veerle Coup (EUR)
Sensitivity Analyis of Decision-Theoretic Networks

2000-9 Florian Waas (CWI)
Principles of Probabilistic Query Optimization

2000-10 Niels Nes (CWI)
Image Database Management System Design Considerations, Algorithms and Architecture

2000-11 Jonas Karlsson (CWI)
Scalable Distributed Data Structures for Database Management

2001-1 Silja Renooij (UU)
Qualitative Approaches to Quantifying Probabilistic Networks

2001-2 Koen Hindriks (UU)
Agent Programming Languages: Programming with Mental Models

2001-3 Maarten van Someren (UvA)
Learning as problem solving

2001-4 Evgueni Smirnov (UM)
Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets

2001-5 Jacco van Ossenbruggen (VU)
Processing Structured Hypermedia: A Matter of Style

2001-6 Martijn van Welie (VU)
Task-based User Interface Design

2001-7 Bastiaan Schonhage (VU)
Diva: Architectural Perspectives on Information Visualization

2001-8 Pascal van Eck (VU)
A Compositional Semantic Structure for Multi-Agent Systems Dynamics

2001-9 Pieter Jan 't Hoen (RUL)
Towards Distributed Development of Large Object-Oriented Models, Views of Packages as
Classes

2001-10 Maarten Sierhuis (UvA)
Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language
for work practice analysis and design

2001-11 Tom M. van Engers (VUA)
Knowledge Management: The Role of Mental Models in Business Systems Design

2002-01 Nico Lassing (VU)
Architecture-Level Modifiability Analysis

2002-02 Roelof van Zwol (UT)
Modelling and searching web-based document collections

2002-03 Henk Ernst Blok (UT)
Database Optimization Aspects for Information Retrieval

2002-04 Juan Roberto Castelo Valdueza (UU)
The Discrete Acyclic Digraph Markov Model in Data Mining

2002-05 Radu Serban (VU)
The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned
Agents

2002-06 Laurens Mommers (UL)
Applied legal epistemology; Building a knowledge-based ontology of the legal domain

2002-07 Peter Boncz (CWI)
Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications

2002-08 Jaap Gordijn (VU)
Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas

2002-09 Willem-Jan van den Heuvel (KUB)
Integrating Modern Business Applications with Objectified Legacy Systems

2002-10 Brian Sheppard (UM)
Towards Perfect Play of Scrabble

2002-11 Wouter C.A. Wijngaards (VU)
Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12 Albrecht Schmidt (Uva)
Processing XML in Database Systems

2002-13 Hongjing Wu (TUE)
A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU)
Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT)
Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16 Pieter van Langen (VU)
The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UVA)
Understanding, Modeling, and Improving Main-Memory Database Performance

2003-01 Heiner Stuckenschmidt (VU)
Ontology-Based Information Sharing in Weakly Structured Environments

2003-02 Jan Broersen (VU)
Modal Action Logics for Reasoning About Reactive Systems

2003-03 Martijn Schuemie (TUD)
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04 Milan Petkovic (UT)
Content-Based Video Retrieval Supported by Database Technology

2003-05 Jos Lehmann (UVA)
Causation in Artificial Intelligence and Law - A modelling approach

2003-06 Boris van Schooten (UT)
Development and specification of virtual environments

2003-07 Machiel Jansen (UvA)
Formal Explorations of Knowledge Intensive Tasks

2003-08 Yongping Ran (UM)
Repair Based Scheduling

2003-09 Rens Kortmann (UM)
The resolution of visually guided behaviour

2003-10 Andreas Lincke (UvT)
Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture

2003-11 Simon Keizer (UT)
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT)
Dutch speech recognition in multimedia information retrieval

2003-13 Jeroen Donkers (UM)
Nosce Hostem - Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN)
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD)
Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI)
Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17 David Jansen (UT)
Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM)
Learning Search Decisions

2004-01 Virginia Dignum (UU)
A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02 Lai Xu (UvT)
Monitoring Multi-party Contracts for E-business

2004-03 Perry Groot (VU)
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04 Chris van Aart (UVA)
Organizational Principles for Multi-Agent Architectures

2004-05 Viara Popova (EUR)
Knowledge discovery and monotonicity

2004-06 Bart-Jan Hommes (TUD)
The Evaluation of Business Process Modeling Techniques

2004-07 Elise Boltjes (UM)
Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral
voor meisjes

2004-08 Joop Verbeek (UM)
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële
gegevensuitwisseling en digitale expertise

2004-09 Martin Caminada (VU)
For the Sake of the Argument; explorations into argument-based reasoning

2004-10 Suzanne Kabel (UVA)
Knowledge-rich indexing of learning-objects

2004-11 Michel Klein (VU)
Change Management for Distributed Ontologies

2004-12 The Duy Bui (UT)
Creating emotions and facial expressions for embodied agents

2004-13 Wojciech Jamroga (UT)
Using Multiple Models of Reality: On Agents who Know how to Play

2004-14 Paul Harrenstein (UU)
Logic in Conflict. Logical Explorations in Strategic Equilibrium

2004-15 Arno Knobbe (UU)
Multi-Relational Data Mining

2004-16 Federico Divina (VU)
Hybrid Genetic Relational Search for Inductive Learning

2004-17 Mark Winands (UM)
Informed Search in Complex Games

2004-18 Vania Bessa Machado (UvA)
Supporting the Construction of Qualitative Knowledge Models

2004-19 Thijs Westerveld (UT)
Using generative probabilistic models for multimedia retrieval

2004-20 Madelon Evers (Nyenrode)
Learning from Design: facilitating multidisciplinary design teams

2005-01 Floor Verdenius (UVA)
Methodological Aspects of Designing Induction-Based Applications

2005-02 Erik van der Werf (UM))
AI techniques for the game of Go

2005-03 Franc Grootjen (RUN)
A Pragmatic Approach to the Conceptualisation of Language

2005-04 Nirvana Meratnia (UT)
Towards Database Support for Moving Object data

2005-05 Gabriel Infante-Lopez (UVA)
Two-Level Probabilistic Grammars for Natural Language Parsing

2005-06 Pieter Spronck (UM)
Adaptive Game AI

2005-07 Flavius Frasincar (TUE)
Hypermedia Presentation Generation for Semantic Web Information Systems

2005-08 Richard Vdovjak (TUE)
A Model-driven Approach for Building Distributed Ontology-based Web Applications

2005-09 Jeen Broekstra (VU)
Storage, Querying and Inferencing for Semantic Web Languages

2005-10 Anders Bouwer (UVA)
Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments

2005-11 Elth Ogston (VU)
Agent Based Matchmaking and Clustering - A Decentralized Approach to Search

2005-12 Csaba Boer (EUR)
Distributed Simulation in Industry

2005-13 Fred Hamburg (UL)
Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen

2005-14 Borys Omelayenko (VU)
Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics

2005-15 Tibor Bosse (VU)
Analysis of the Dynamics of Cognitive Processes

2005-16 Joris Graaumans (UU)
Usability of XML Query Languages

2005-17 Boris Shishkov (TUD)
Software Specification Based on Reusable Business Components

2005-18 Danielle Sent (UU)
Test-selection strategies for probabilistic networks

2005-19 Michel van Dartel (UM)
Situated Representation

2005-20 Cristina Coteanu (UL)
Cyber Consumer Law, State of the Art and Perspectives

2005-21 Wijnand Derks (UT)
Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics

2006-01 Samuil Angelov (TUE)
Foundations of B2B Electronic Contracting

2006-02 Cristina Chisalita (VU)
Contextual issues in the design and use of information technology in organizations

2006-03 Noor Christoph (UVA)
The role of metacognitive skills in learning to solve problems

2006-04 Marta Sabou (VU)
Building Web Service Ontologies

2006-05 Cees Pierik (UU)
Validation Techniques for Object-Oriented Proof Outlines

2006-06 Ziv Baida (VU)
Software-aided Service Bundling - Intelligent Methods and Tools for Graphical Service Modeling

2006-07 Marko Smiljanic (UT)
XML schema matching – balancing efficiency and effectiveness by means of clustering

2006-08 Eelco Herder (UT)
Forward, Back and Home Again - Analyzing User Behavior on the Web

2006-09 Mohamed Wahdan (UM)
Automatic Formulation of the Auditor’s Opinion

2006-10	Ronny Siebes (VU) Semantic Routing in Peer-to-Peer Systems
2006-11	Joeri van Ruth (UT) Flattening Queries over Nested Data Types
2006-12	Bert Bongers (VU) Interactivation - Towards an e-cology of people, our technological environment, and the arts
2006-13	Henk-Jan Lebbink (UU) Dialogue and Decision Games for Information Exchanging Agents
2006-14	Johan Hoorn (VU) Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change
2006-15	Rainer Malik (UU) CONAN: Text Mining in the Biomedical Domain
2006-16	Carsten Riggelsen (UU) Approximation Methods for Efficient Learning of Bayesian Networks
2006-17	Stacey Nagata (UU) User Assistance for Multitasking with Interruptions on a Mobile Device
2006-18	Valentin Zhizhkun (UVA) Graph transformation for Natural Language Processing
2006-19	Birna van Riemsdijk (UU) Cognitive Agent Programming: A Semantic Approach
2006-20	Marina Velikova (UvT) Monotone models for prediction in data mining
2006-21	Bas van Gils (RUN) Aptness on the Web
2006-22	Paul de Vrieze (RUN) Fundaments of Adaptive Personalisation
2006-23	Ion Juvina (UU) Development of Cognitive Model for Navigating on the Web