ARTICLE IN PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Int. J. Human-Computer Studies ▮ (▮▮▮▮) ▮▮▮–▮▮▮

International Journal of Human-Computer Studies

www.elsevier.com/locate/ijhcs

Classification of user image descriptions

L. Hollink[a,*], A.Th. Schreiber[a], B.J. Wielinga[b], M. Worring[c]

[a] *Free University of Amsterdam, Business Informatics, De Boelelaan 1081a, NL-1081 HV Amsterdam, Netherlands*

[b] *University of Amsterdam, Social Science Informatics, Roetersstraat 15, NL-1018 WB Amsterdam, Netherlands*

[c] *University of Amsterdam, Intelligent Sensory Information Systems, Kruislaan 403, NL-1098 SJ Amsterdam, Netherlands*

Received 3 February 2003; accepted 23 March 2004

Abstract

In order to resolve the mismatch between user needs and current image retrieval techniques, we conducted a study to get more information about what users look for in images. First, we developed a framework for the classification of image descriptions by users, based on various classification methods from the literature. The classification framework distinguishes three related viewpoints on images, namely nonvisual metadata, perceptual descriptions and conceptual descriptions. For every viewpoint a set of descriptive classes and relations is specified. We used the framework in an empirical study, in which image descriptions were formulated by 30 participants. The resulting descriptions were split into fragments and categorized in the framework. The results suggest that users prefer general descriptions as opposed to specific or abstract descriptions. Frequently used categories were objects, events and relations between objects in the image.

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.

*Corresponding author. Tel.: +31-20-444-7740; fax: +31-20-444-7728.

E-mail addresses: hollink@swi.psy.uva.nl, hollink@cs.vu.nl (L. Hollink), schreiber@cs.vu.nl (A.Th. Schreiber), wielinga@swi.psy.uva.nl (B.J. Wielinga), worring@science.uva.nl (M. Worring).

1071-5819/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2004.03.002

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. Keyword-based retrieval relies on textual descriptions accompanying an image. A disadvantage of this 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 such as color, shape and texture, 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 recognized the lack of knowledge about the way users search for images ((Choi and Rasmussen, 2001; Fidel, 1997), among others). The aim of this paper is to investigate user needs in image retrieval. We do this by asking two questions: (1) which categories of image descriptions exist?, and (2) how much 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.

We used this framework 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 categorized in the framework. The results show how much each category is used in various category search tasks.

2. Related work

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

2.1. Theories

Erwin Panofsky developed a theory to structure content descriptions of images as early as 1962 (Panofsky, 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." 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 categorized 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 image descriptions. The so-called Panofsky/Shatford model has become a widespread 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, color and number of colors. The next three perceptual levels are based on the low-level features "color," "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 Panofski/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 *indexes* 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. Empirical studies

Although little empirical knowledge about this topic is available, some experiments from various fields have added useful information to the theories above. Enser and McGregor (1992) analyzed user requests submitted to the Hulton Deutch CD Collection. They represented the requests in terms of a 2×2 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 analyzed free text image descriptions and deduced 12 classes of image attributes that were used in the descriptions. She found that people describe images mainly using "object," "people," "color" and "story" classes. The latter contains descriptions of the event, the setting, activities and time aspects. The 12 classes of Jörgensen and the 2×2 matrix of Enser and McGregor are 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 from 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 botanical 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.

2.3. Practical utilization

Practical improvements in the field of indexing and searching have been accomplished through the introduction of the Dublin Core Metadata standard (http://dublincore.org). 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 focusing only on images, are the Visual Resources Association (VRA) Core Categories. The VRA (http://www.VRAweb.org) formulated an "element set to create records to describe works of visual culture as well as the images that document them" (Visual resource association data standards committee, 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 content element, which allows free text descriptions. 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 3.2.

3. A framework for the classification of image descriptions

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 between 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 categorizes image descriptions.

Not only the *focus*, but also the *form* of the discussed classification methods varies. Jörgensen, for example, categorizes descriptions into twelve topics, while Jaimes and Chang use 10 successive levels. Variations in form make it difficult to see the differences and similarities between categories. To cope with this, we use UML to visualize the framework. The Unified Modeling Language (UML) (Booch et al., 1998) is a well-defined, standardized modeling language.

To realize 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 add to this an additional category: the nonvisual information. This results in the three toplevels of the framework: the nonvisual level, the perceptual level and the conceptual level (Fig. 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 *nonvisual* level are a subset of the VRA element set. The elements 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 colors and shape. At the *conceptual* level the content of the image is described. The

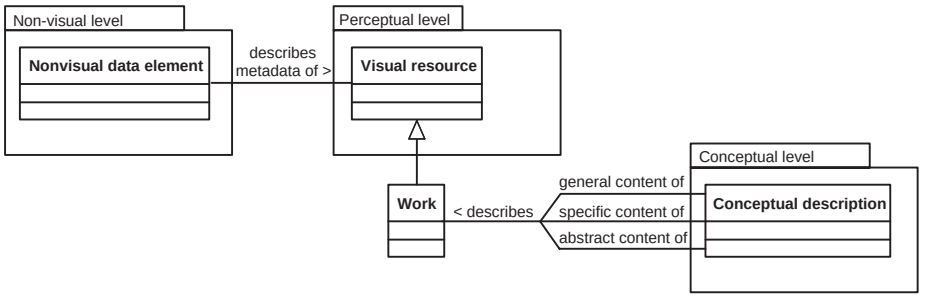

Fig. 1. UML package diagram of an integrated framework for the classification of image descriptions.

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 visual resource association (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, the International Federation of Library Associations and Institutions (IFLA study group on the functional requirements for bibliographic records, 1998), 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 "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.

From Fig. 1 it can be seen that the nonvisual and perceptual levels give information about visual resources. The resources are not necessarily works. The nonvisual class **date** or the perceptual class **color** 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 nonvisual, perceptual and conceptual levels in detail in the following sections.

3.2. Nonvisual level

At the nonvisual level we are interested in descriptive information about the image. The information at this level is often called metadata. Hillmann (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.

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

Fig. 2 shows the nonvisual data elements 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 elements speak for themselves.

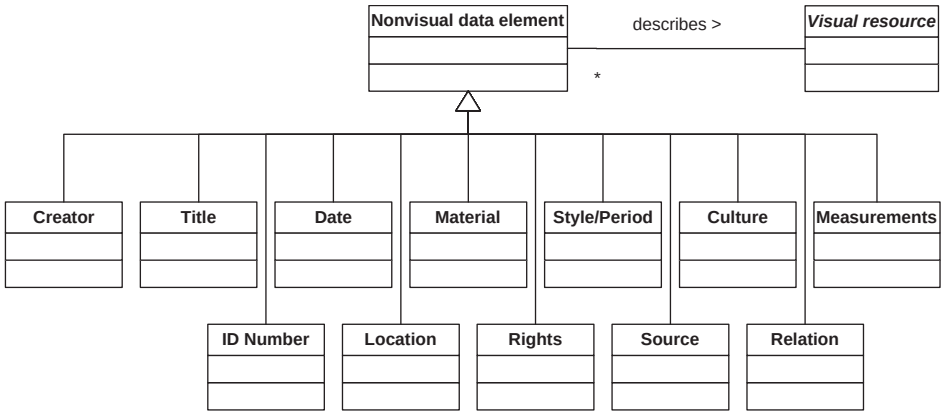

Fig. 2. UML class diagram of the nonvisual level.

The elements are a subset of VRA elements that meet the criteria for the nonvisual 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.

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 color and visual element classes in Jörgensen (1998).

The structure of the perceptual level is visualized in a UML class diagram (Fig. 3) The **visual resource** class is in the center 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:

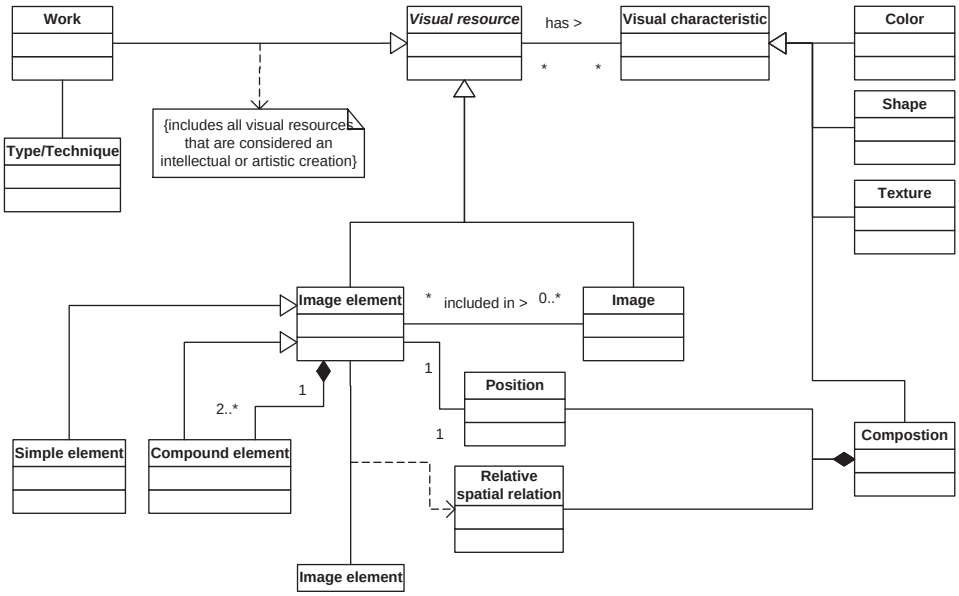

Fig. 3. UML class diagram of the perceptual level.

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 **color**, **shape** and **texture**. At this point, the relation between image descriptions and the 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.

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, p. 290): "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 basic level of abstraction at which [one] can obtain the most information with the least cognitive effort" (Rosch, 1973). 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 visualized in Fig. 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 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.

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 **color** 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.

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.

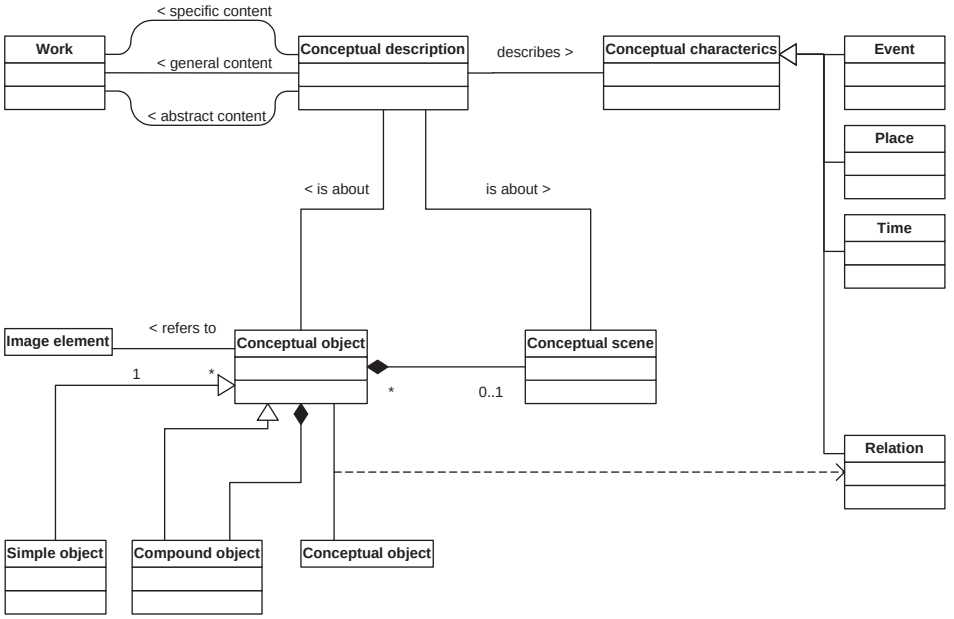

Fig. 4. UML class diagram of the conceptual level.

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 expertise of the user and (3) the task the user is performing. An overview of the user characteristics is shown in Fig. 5.

3.5.1. 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

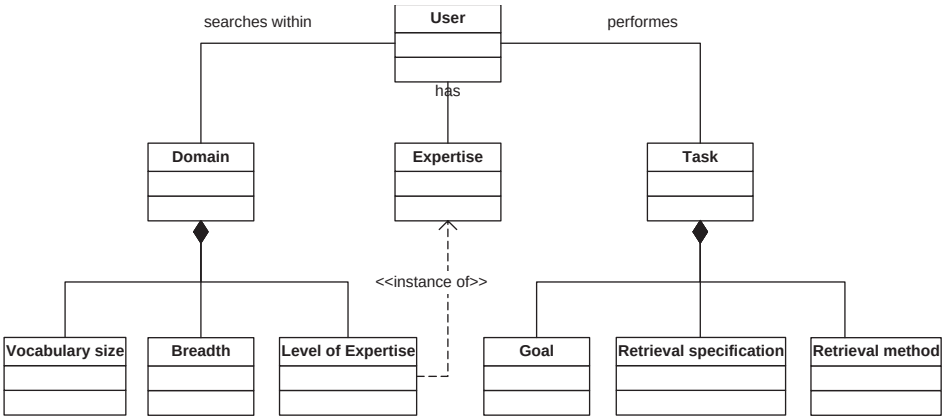

Fig. 5. Classification of users.

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 specialized 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.

3.5.2. 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) analyzed 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 use of broader terms decreased. Frost and Noakes (1998) demonstrate 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.

3.5.3. 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. The way an image will be used affects how a person searches for that image. A goal could be to decorate the website of a beach resort by adding pictures of people having fun, to learn the structure of DNA by looking at an image of a double helix, or to proof something by showing its existence on an image. Fidel (1997) organizes 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. 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. A query for an image of a DNA double helix to learn its structure is also at the data pole. At the objects pole images are treated as objects. They may be used as decoration or to represent ideas. The illustration of a website is an example of a search at the objects pole.

Retrieval specification: The retrieval specification is the set of conditions the user expresses as input for the search. The conditions can be more or less precise and 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 Smeulders et al. (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.

User characteristics are a research topic on their own. Studies like the ones by Vakkari (2000) and Frost and Noakes (1998) show the relations between user related 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.

4. User study

4.1. Aim of the study

In Section 1, the 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 2.1, 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 (see Section 4.2), are both category search tasks. Category search has been described in Section 3 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" and a "query" task?
3. What is the difference in the use of classes by participants in different domains?

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. First, they wrote down a free text description of the

image. Second, 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 complex and time-consuming 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" 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, while the third was a paragraph in a news item in a Dutch newspaper. The characteristics of these domains are not further specified in this paper.

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]

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 in the study, we compared the results of the three texts. Our aim is to get an idea of the influence of the domain on the use of search terms. The differences between the texts suggest some differences in the results. The

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

second text, which contains descriptions of colors 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 is different from approaches in previous experiments (Heidorn, 1999; Jörgensen, 1998) in that the participants are 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 seeing 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.

In the following examples of descriptions and queries the use of various classes can be seen. 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 utilization of the abstract level in the phrase "narrow-minded dullness." All three 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 generalizations of words in the texts. Copied fragments consisted mainly of general objects from Text 1 (15% of all descriptions), general objects from Text 2 (5% of all descriptions) and specific

objects from text 3 (8% of all descriptions). The copied fragments were nevertheless included in the analysis for two reasons. First, the participants picked these words from the text since they considered them to be relevant for their imagined image. Second, 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.

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.

4.4. Analysis of the data

After collection of the descriptions, they were first split into fragments suitable for categorization. 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 recognizable, | birds | fly | in the air. | The image | beams much warmth | because of the dark pink color | 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 normalized 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 and $N(d_i)$ is the total number of fragments in a description i. 180 Is the total number of descriptions as given by 30 participants performing two tasks on three domains. Finally, the results were analyzed 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 synchronize the classification decisions we used a set of guidelines for each part of the classification. The most important ones are the following:[2]

1. Consider a fragment to be specific only if it is more specific than Rosch's basic level categories, as explained in Section 3.4 (Rosch, 1973).
2. Categorize 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[3] was used as a measure of agreement between reviewers. We found a correspondence of 70% (Cohen's kappa = 0.66) with the first additional reviewer and 75% (Cohen's kappa = 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 (nonvisual, 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 color 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 (Cohen's kappa = 0.66) and 85% for the second reviewer (Cohen's kappa = 0.67). The second part corresponded in 88% (Cohen's kappa = 0.68) and 90% (Cohen's kappa = 0.74) of the classifications. The third part reached a correspondence of 88% (Cohen's kappa = 0.83) and 89% (Cohen's kappa = 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

[2] A complete copy of the guidelines can be found at http://www.cs.vu.nl/~laurah/guidelines.pdf

[3] Cohen's Kappa measures concordance between 2 raters using nominal data. Kappa varies between −1.0 and 1.0. The degree of concordance is considered sufficient if kappa is larger than 0.60 (Brink et al., 2002).

due to three classes. First, the difference between the perceptual composition class and the conceptual relation class caused dissimilarity. Second, opinions about whether an object or scene could be specified as a place differed. Third, 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 Nonunique categories.

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, while 12% were perceptual and only 0.9% nonvisual (Table 1). The near absence of nonvisual 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. Thirty-two 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 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 4.4). The absolute numbers do therefore not always correspond to the percentages.

In Section 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 5.2, we look at the differences in element occurrence between a "describe" and a "query" task, and in Section 5.3, we discuss the differences between the three domains.

5.1. Use of classes of image descriptions in a category search task

5.1.1. 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: color, shape, texture, composition and technique. Within the perceptual level, the composition class was used most: 37.1% (Table 2). These were mainly terms describing the relative spatial relationships between the elements.

Table 1
Levels of abstraction in image descriptions in absolute numbers of occurrences and weighted percentages

Level	Count	%
Nonvisual	7	0.9
Perceptual	184	11.9
Conceptual	928	87.2
Total	1119	100.0

Table 2
Perceptual level: scope and characteristics of image descriptions in absolute numbers of occurrences and weighted percentages

Scope	Object		Scene		Total	
Characteristic	Count	%	Count	%	Count	%
Color	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

Table 3
Conceptual level: occurrences of conceptual classes in absolute numbers and weighted percentages

Level		**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
	Uncharacterized	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
	Uncharacterized	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

Examples are "an object (dragonfly) *above* an object (pool)" or "an object (squirrel) *on* an object (the back of a dragonfly)." Color accounts for 32.1% of the perceptual elements and technique for 28.4%. Shape was hardly used (2.5%) 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, 54.4% and 45.6%, respectively.

5.1.2. Conceptual levels

The general sublevel was the most frequently used level within the conceptual level (74.4% of all conceptual descriptions). The remaining two sublevels, the specific sublevel and the abstract sublevel, were used less, 16.4% and 9.2%, respectively (Table 3). At the conceptual levels, objects were used more than twice as much as scenes (70.3% and 29.7%).

The descriptions were also categorized 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: 57.1% of all conceptual fragments. Out of the four characteristics, event was used most (13.3%), followed by place (12.5%), time (6.2%) and relation (2.1%) (Table 3).

A log-linear analysis[4] was undertaken to test the existence of a relationship between the components 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 $3 \times 2 \times 5$ contingency table (Table 3). The analysis fitted a model to the three variables that best explains the frequency data in the table:

[*SublevelScope*][*ScopeCharacteristic*].

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 ($X^2 = 13.2$, $df = 16$, $p = 0.34$). 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 (e.g. "morning," "in spring" or "may 15") are always associated with a scene, while on average descriptions are more often about objects.

Further research is needed to test these hypotheses.

5.2. Differences between a "describe" 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 (Tables 4 and 5). Differences can be seen at the remaining levels: in the "query"

[4] Log-linear analysis is a version of chi-square analysis that is used to analyze 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 terms: expected value = constant + (row term) + (column term) + (association term). An alternative representation of this model is the shorter notation $[AB]$. X^2 values are computed for all possible models to test which model fits the data best (Wickens, 1989).

Table 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	%	Count	%
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

Table 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

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 et al., 2002) showed that in our study there is indeed a significant difference in the distribution of descriptions over (sub)levels in the two tasks ($X^2 = 10.6$, 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. We are interested in further research to test this hypothesis. 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.

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. Descriptions of illustrations for the first text, a paragraph from a children's book, contain less abstract fragments than average (3.2%). The third text, a newspaper item, contains more specific descriptions than average (36.4%). 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 (41.3%). 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 ($X^2 = 39.6$, 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 of more scene descriptions than object descriptions (53.0% and 46.9%, respectively). A chi-square test showed a significant difference at this point ($X^2 = 17.1$, df $= 2$, $p = 0.00$). No indication has been 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 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.

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 two times 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 has been 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 nonvisual 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 are not biased by the limitations of existing search interfaces, while the queries in the "query" task are.

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 11–14% of all descriptions. 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 specializations 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.

It is difficult to compare results across experiments. This is due to the different use of classes to categorize 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: object and people are the largest of the 12 classes (Jörgensen, 1998) and together account for 39%, while none of the other classes exceeds the 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 Jörgensens 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 Panofski/Shatford model (Shatford, 1986). The **object** class in our framework is the equivalent of the *who* question in the Panofski/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 our own can be found in the use of the specific level. In the study of Armitage and Enser (1997) the specific level is the most frequently used level, while in the present study the general level is by far the most 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 in the newspaper domain than in the novel and children's domains. 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 paper was to contribute to the knowledge about one side of the gap, the concepts of users. The next step will be 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 experiment show that the conceptual object class is the class that users typically use in image descriptions. A perceptual element is described by color, 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.

References

Armitage, L., Enser, P., 1997. Analysis of user need in image archives. Journal of Information Science 23 (4), 287–299.

Booch, G., Rumbaugh, J., Jacobson, I., 1998. The Unified Modeling Language User Guide. Addison-Wesley, Reading, MA.

Brink, W.P., Van den, Koele, P., 2002. Statistiek, Vol. 3. Boom.

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., Rasmussen, E., 2001. Users' relevance criteria in image retrieval in American history. Information Processing and Management 38, 695–726.

Eakins, J., 1998. Techniques for image retrieval. Library and Information Briefings 85. Library Information Technology Centre, South Bank University, London.

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

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

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., Yanker, P., 1995. Query by image and video content: the qbic system. IEEE Computer 28 (9), 23–32.

Frost, C.O., Noakes, A., 1998. Browsing images using broad classification categories. In: 9th ASIS SIGCR Classification Research Workshop, October 25, Pittsburg, PA, pp. 71–79.

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

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

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

Heidorn, P., 1999. The identification of index terms in natural language object descriptions. In: Proceedings of the American Society for Information Science Conference.

Hillmann, D., 2001. Using Dublin core. Recommendation, Dublin Core Metadata Initiative.

IFLA study group on the functional requirements for bibliographic records, 1998. Functional requirements for bibliographic records. Final report, International Federation of Library Associations and Institutions, Munchen.

Jaimes, A., Chang, S.-F., 2000. A conceptual framework for indexing visual information at multiple levels. In: SPIE Internet Imaging 2000, Vol. 3964.

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

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

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, Vol. 3964. ACM Press, NY, pp. 212–218.

Panofsky, E., 1962. Studies in Iconology. Harper and Row, New York.

Rosch, E., 1973. Natural categories. Cognitive Psychology 4, 328–350.

Schreiber, A., Dubbeldam, B., Wielemaker, J., Wielinga, B., 2001. Ontology-based photo annotation. IEEE Intelligent Systems 16 (3), 66–74.

Shatford, S., 1986. Analyzing the subject of a picture: a theoretical approach. Cataloging and Classification Quarterly 6 (3), 39–62.

Smeulders, A., Worring, M., Santini, S., Gupta, A., 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.

Vakkari, P., 2000. Cognition and changes of search terms and tactics during task performance; a longitudinal study. In: Proceedings of the RIAO 2000 Conference Paris, pp. 894–907.

Visual resource association data standards committee, 2002. VRA core categories, version 3.0. Technical report, Visual Resources Association.

Wickens, T., 1989. Multiway Contingency Tables Analysis for the Social Sciences. Lawrence Erlbaum Associates, Hillsdale, NJ, 07642.