The image retrieval task: implications for the design and evaluation of image databases

Raya Fidel
Graduate School of Library and Information Science, University of Washington, Seattle, WA 98195

A review of studies about searching behaviour in image retrieval suggests that retrieval tasks may affect searching behaviour. Retrieval tasks occur along a spectrum starting with the Data Pole, which involves retrieval of images for the information which the images include, and ending with the Objects Pole, which concerns the retrieval of images as objects. Each Pole generates a certain searching behaviour which has characteristics opposing those of the other Pole. These characteristics suggest that: (a) Relevance feedback may not be useful for tasks on the Objects Pole; (b) Measuring precision on the Data Pole should be replaced with another measurement of effort and time, while on the Objects Pole, the quality of browsing sets and the precision of the browsing process should be measured instead of precision; and (c) Recall is not useful for the Data Pole, and requires much exploration before it can be adopted for the Object Pole. Additional research in searching behaviour and about performance measurement will improve retrieval from image databases.

With increased access to digitized images, the interest in image retrieval has soared. Much research is focused on image indexing and retrieval – mostly on the technical aspects. Investigators are experimenting with image retrieval using a variety of approaches with algorithms (e.g.,[1, 2]) and AI techniques (e.g.,[3]). Some approaches are based on the images themselves (content-based retrieval) and others on the text in images or around images.

Notwithstanding these research activities, a number of fundamental issues in image retrieval have been somewhat ignored. For example:

- What are the differences between image and text retrieval?
- What image attributes are important for retrieval? For instance, are the commonly used attributes of colour, shape and texture useful?
- What are the characteristics of users' queries for images?

Two among these fundamental issues have been repeatedly mentioned as central concerns (e.g.,[4, 5]):

- The limited research in user issues, and
- The lack of theoretical background for the design and evaluation of image databases.

These two issues are related because it is desirable to base the design and evaluation of image databases on an understanding of image seeking and searching behaviour of users. Few studies have attempted to analyze this behaviour, and it is not surprising that very little progress has been made in developing methods and standards to design and evaluate image databases. This paper examines the few studies on searching behaviour in image retrieval and begins to explore some of their implications for the design and evaluation of image databases.

Conceptual considerations

Various attributes differentiate images from text. Enser[5] explained that while it is more difficult to gain acess to text documents – and even more so to retrieve only relevant ones – than to generate such documents, the distinction between ease of generation and difficulty of access and retrieval is much more marked for images. He also reported that various experiments in image retrieval had assumed that relevance assessment for images could be done much more rapidly than for text.

General standards for the creation of metadata, such as the Dublin Core[15] and the Anglo-American Cataloging Rules,[16] have another approach altogether. They aim at establishing metadata elements that are useful for the retrieval of *both* text and images. Rather than highlighting the unique features of image retrieval, their goal is to provide a basic and universal set of elements that constitute metadata for all resources or works.

Layne[6] examined image attributes that are important for indexing and retrieval. While she did not conduct a systematic user study, she derived these attributes by integrating her long experience as an art librarian with theoretical considerations. Four facets play a role in image indexing and retrieval, but only one of them, subject attributes, is usually considered in text indexing:

1. Biographical attributes. These relate to the 'biography' of an image and are of two types: (a) Those related to the **creation** of an image, such as the name of its creator, the time and place of creation; and (b) The image **'travel'**, that is, where it is now, who has owned it, where it has been, etc.

2. Subject attributes. Both text and images have subject attributes, but their manifestation in image indexing is richer than in text. Guided by Panofsky's modes of image analysis,[7] Layne explains that an image can be of one thing, and at the same time *about* something else. For example, a picture of a mother and a child is *of* a mother and a child, but it can be *about* motherly love and another such picture *about* the Immaculate Conception. In addition, an image's subject can be general and specific at the same time. The image of a mother and a child can be

perceived as a picture of two human figures, which is broader than 'mother and child,' or as one of Mary and baby Jesus, which is narrower.

3. Exemplified attributes. An image can be an example of something else, such as a picture of a poster.

4. Relationships attributes. An image can have a relationship with another image. For example, a preliminary drawing and the finished painting, or an architectural plan and an image of the finished building.

While these types of attributes are sometimes represented in indexing of text documents, they are much more important for indexing and retrieval of images, particularly if they are works created by artists.

Analysis of search requests

Enser[8] carried out the first analysis of user requests for images in a real-life situation. The environment in which his study took place was different from that of Layne. He analyzed over 2,700 requests that had been submitted to the largest picture archive in Europe, the Hulton Deutsch Collection Limited. Users came primarily from among book publishers, advertising and design companies, and from magazine and newspaper publishers. The analysis was based on the request forms only, and was carried out after the requests had been answered.

In his analysis of the requests, Enser used two characteristics: (a) Whether or not a request was for a unique person, object or event; and (b) Whether or not a request was further refined by the facets time, place, action, event or technical specifications.

The study revealed that almost 70% of the requests were for a unique person, object or event, and that most of the other requests included refinements, mostly by time. In addition, Enser found that only requests for a unique person, object or event that were *not* refined by any other facet could be searched easily by the classification scheme used. The rest required some browsing. He concluded that because the majority of the requests were for a unique person, object or event, and because the classification scheme was not useful for the other requests, a free-text retrieval based on the images' captions should be satisfactory for a general-purpose collection such as the one examined.

In a later study[17] Armitage and Enser analyzed an additional set of over 1,700 requests from seven different libraries. Combining the unique/refined distinction with Panofsky's categories,[7] they presented a general typology of requests that could inform the design of interfaces for end-user searching for images of all types.

Analysis of user behaviour

Unlike Enser's study, two investigations examined seeking and searching behaviour in controlled experiments where subjects interacted

with images that were selected for them by the researchers. Korf Vidal's study[9] was limited to an interaction with 48 images of one object: the Brooklyn Bridge. She asked each of her 58 subjects to sort the images into piles 'in a way that makes personal sense.' She then generated cluster maps of images using the Q-sort method. Results showed that there were categories of images that were common to the majority of the subjects.

The second study, carried out by Jorgensen,[10] employed a variety of images and aimed at naming categories common to users. Jorgensen selected randomly 77 images from *The Society of Illustrators 25th Annual of American Illustration.* Subjects interacted with images in the lab, and were asked to think aloud. She videotaped the interactions, and analyzed the think-aloud protocols to find what attributes the subjects used when they interacted with the images.

The study's subjects completed three tasks:

- **Describing task** in which they viewed six projected images and wrote a description of each (N=48);
- **Searching task** in which subjects were each given two terms representing abstract concepts, such as happy, mysterious, and then browsed the set of 77 images to find those relevant to the queries (N=18); and
- **Sorting task** in which the same subjects sorted the 77 images into groups for their own use as if the images were their personal collection (N=18).

The analysis of the verbal protocols revealed 12 classes of image attributes referred to by the subjects. These are listed in Table 1. Jorgensen also distinguished between *Perceptual* (*P*) and *Interpretive* (*I*) attributes. The value of a perceptual attribute can be determined just by looking at an image. For example, one may point out that the flower in a picture is a rose, and that its color is pink. The value of an interpretive attribute requires some personal reflection and abstraction, and therefore might be in the eyes of the viewer. For example, only an interpretation of an image can lead a viewer to determine that the atmosphere of the image is gloomy, or that it is painted in a romantic style. In addition, she defined *Reactive* attributes as those 'which include mental activity on the part of the participants such as conjecture or emotional response to the pictures.' ([10], p. 125)

Attribute Class	Description
Literal object (*P*)	Named objects which are visually perceived, e.g., body parts, clothing.
People (*P*)	The presence of a human form.
People-related attributes (*I*)	The nature of the relationship among people, social status or emotions.
Art historical information (*I*)	Information related to the production context of the image, e.g., artists, medium, style.
Color (*P*)	Specific named colors or terms relating to various aspects of color.
Visual elements (*P*)	Elements such as composition, focal point, motion, shape, texture.
Location (*P*)	Both general and specific locations within the image.
Description (*P*)	Descriptive adjectives, e.g., wooden, elderly, or size or quantity.
Abstract concepts (*I*)	Attributes such as atmosphere, theme, or symbolic aspects.
Content/story (*I*)	A specific instance being depicted.
External relationships (*I*)	Relationships to attributes within or without the image, e.g., similarity.
Viewer response (*Reactive*)	Personal reaction to the image.

TABLE 1: *Classes of image attributes (after Jorgensen[10]) P=Perceptual; I=Interpretive*

After defining the attribute classes, Jorgensen tallied the occurrences of each class in the performance of each task. The results are presented in tables 2-4. These clearly showed that the prominence of a class depended on the task the subjects were carrying out. While similar classes were used frequently in the *describing* and *searching* tasks, subjects concentrated on other classes during the *sorting task.*

Attribute Class	Percent
Literal object (*P*)	34%
Color (*P*)	9%
People (*P*)	9%
Location (*P*)	8%
Content/story (*I*)	7%
Visual elements (*P*)	7%
Description (*P*)	6%
People-related (*I*)	5%
Art historical (*I*)	4%
Viewer response (*R*)	4%
External relationships (*I*)	3%
Abstract concepts (*I*)	3%

TABLE 2: *Describing task*

Attribute Class	Percent
Literal object (*P*)	27%
Content/story (*I*)	11%
Location (*P*)	11%
People (*P*)	10%
Color (*I*)	10%
Description (*P*)	9%
Art historical (*I*)	6%
Visual elements (*P*)	5%
People-related (*I*)	4%
External relationships (*I*)	4%
Viewer response (*R*)	2%
Abstract concepts (*I*)	1%

TABLE 3: *Searching task*

Attribute Class	Percent
Art historical (*I*)	24%
Viewer response (*R*)	14%
Abstract concepts (*I*)	14%
Literal object (*I*)	9%
External relationships (*I*)	9%
People (*P*)	9%
Content/story (*I*)	8%
People-related (*I*)	4%
Visual elements (*P*)	4%
Description (*P*)	3%
Color (*P*)	3%
Location (*P*)	1%

TABLE 4: *Sorting task*

Another analysis of search requests

The classes of attributes that Jorgensen uncovered, and the finding that category use may depend on the task in which a user is involved, suggested a new way to examine image retrieval. In particular, it raised the question: Should the design and evaluation of image databases be guided by the tasks involved in image retrieval? Even more specifically: Should we look for performance measurements that apply to all retrieval tasks, or does each task require its own measurement?

To begin to answer these questions, we conducted a small exploratory study in which we analyzed 100 actual requests using Jorgensen's attribute classes. The agency we selected had a large collection of stock photos and a customer base very similar to the one in Enser's study. The agency had a well-developed, in-house controlled vocabulary, and indexing was done by professional indexers. Queries were submitted through a number of channels (phone, fax, mail and email) and the agency's professional staff filled out request forms and

performed the searches. The request forms recorded the description of the query, whether a horizontal or vertical image was desired, whether black-and-white or color, and the number of images the customer was willing to look at. The results of each search was a set of low-resolution images from which the user selected the desired image(s).

Even in this small sample, requests varied in their levels of specificity and abstraction. Some examples are presented in Figure 1.

1. One or more monks meditating in lotus pose. Background optional.
2. Volcanoes: spewing with lava and smoke from top and sides.
3. Fiber optic cable. Cables, general. Anything to convey sending information over the wires. It is for an Internet-related product.
4. Photo of Wassily Kandinsky, or photo of one of his works if first choice not available.
5. Details of cars from the 50's, early 60's & current. None should be white or light colored. Would like to see lots of chrome on the early ones.
6. Emission control/air-traffic control images.
7. Ruins of Japanese battlegrounds.
8. Any Asian artifact from Perry's 1853 journey to Japan.
9. Satellite dishes; big satellite dishes in a row, out in the desert, etc.
10. Close-up of red tropical soil.
11. Mother with child or children. Grandmother with child or children. They can be doing something together, playing, baking, whatever illustrates mother and child interaction. Can be a pregnant mom. No nudity or birth shots acceptable. No Baroque, too modern or too contemporary, or loose impressionist art. Victorian or more detailed impressionist is good.

FIG. 1: *Examples of actual requests*

To 'index' the sample requests with Jorgensen's classes, we assigned all the classes that applied to each request. To do so, we added another class: Object-related attributes, as was required for requests 2 and 9, for example. These describe the relationships between objects or object-related attributes that may evoke emotions, and are interpretive attributes. Apparently, Jorgensen's subjects did not use such attributes. We also observed that it was not always a straightforward matter to determine whether or not a request was abstract because sometimes terms that are concrete actually represent abstract concepts (e.g., requests 3 and 6 in Figure 1). In addition, we noted that among the 100 requests, none asked for a certain shape nor texture, and only two (5 and 10) referred to a certain color.

A tally of the occurrences of each class in the sample requests is presented in Table 5. A comparison with Jorgensen's results (Tables 2-4) showed that the distribution of classes for the *sample requests* was different from that for the other tasks, but was most similar to the *searching task* when considering the top two categories.

Attribute Class	Percent
Literal object (*P*)	27%
Content/story (*I*)	16%
Object-related (*I*)	15%
People (*P*)	13%
People-related (*I*)	13%
Visual elements (*P*)	7%
Art historical (*I*)	4%
Abstract concepts (*I*)	3%
Color (*P*)	2%
Location (*P*)	0%
Description (*P*)	0%
External relationships (*I*)	0%
Viewer response (*R*)	0%

TABLE 5: *Sample requests*

Guided by the assumption that classes that represent interpretive attributes are more difficult to assign in indexing (the *describing task*) than in searching (the *searching task*, and the *sample requests*), we noted that 78% of the *sample requests* included one or more interpretive attribute. To assess the prevalence of such attributes in each task, we collapsed the interpretive classes on the one hand, and the perceptive on the other. Results for all the four tasks are presented in Table 6.

Task	Interpretive	Perceptive	Reactive
Describing	22%	74%	4%
Searching	26%	72%	2%
Sorting	58%	29%	13%
Sample requests	51%	49%	0%

TABLE 6: *Distribution of interpretive and perceptive attributes among tasks*

These results clearly showed the marked difference between the classes used in the *sorting task* and in the *sample requests* on the one hand, and in the *describing* and *searching* tasks on the other. Over half of the attributes used in the *sorting task* and in the *sample requests* were interpretive, while less than a quarter of the attributes for the other tasks were.

One way to explain these puzzling findings is to examine the similarity in results for the *sorting task* and the *sample requests*. Both tasks had the same retrieval task: to find images for personal use. On the other hand, the subjects' assignment in the *describing task* was to describe images, not to retrieve them, and in the *searching task*, the subjects' responsibility was to find images that may satisfy abstract

requests that were given to them. They might have felt uncomfortable making interpretive decisions.

This explanation brings to light an additional factor that the general task may affect searching behaviour: the *nature of the retrieval task*. In other words, what is the nature of the retrieval the user expects? Or, what does the user plan to do with the search results?

The nature of the retrieval task: from data to objects retrieval

Unlike text, it is easy to see that images can be used in various ways. For example, the use of an icon that indicates wheelchair accessibility is very different from the use of a colorful picture waiting to be hung on a wall. The icon is a source of data, or information, while the picture is an object. Thus, images can be used both as sources of data and as objects. What are the differences between tasks that require data retrieval and those that require object retrieval? Should these differences be considered in the design and evaluation of image databases?

Data retrieval tasks and object retrieval tasks are not presented here as a dichotomy but rather as extreme poles in a continuum of retrieval tasks with varying degrees of data and objects as desired retrieval results. At the Data Pole, images are used as sources of information, while at the Objects Pole, images are needed as objects. Let us examine some examples on each pole, and some in-between.

The Data Pole. Systems for the retrieval of cartographic material, medical slides, or chemical structures are examples of systems that store images which are commonly retrieved for the information they embody. A user may want to retrieve a map to see how to get from one place to another; a physician may need a slide of a normal foot to help decide if a patient's foot is flat; and a chemist may use a diagram of a chemical structure to examine the molecular structure of the elements involved. It makes no difference to these users who created the image, its history, or how it relates to other images—as long as the image provides them the information they need.

The Objects Pole. Stock photo agencies are a clear example of organizations that provide retrieval of images as objects. As the examples in Figure 1 illustrate, such agencies may be asked for concrete or abstract images, for a very specific kind of image of a person or an event, or for any image that represents a specific idea or object. What is common to all these requests is the future use of the retrieved images: They all will be used as objects in the products of the requesters, whether as pictures in a history book, as part of an advertisement about the Internet, or on the cover of the next issue of a magazine.

In-Between. Graphic artists, medical instructors, and art historians are examples of users who may retrieve images both as information sources and objects. A graphic artist may want to retrieve

pictures of various trees so he can copy some of them in his next designs, and also to explore the variety of tree shapes. The artist will use the information in the images of trees, as well as the images as objects, to create new images of trees. A medical instructor may look for a good slide of a normal foot for a class she teaches. She wants the slide to have the information required about a normal foot, but, at the same time, she is looking for the best slide as an object: the one that seems clearest to her, the one taken from a useful angle, or with an image big enough to be projected in a classroom. Similarly, when preparing to write an article about the relationship between artists and their cats during the 17th Century, an art historian may want to view all drawings of cats done during that period. The historian uses the images as the raw data to make inferences about the relationship between artists and their cats, but to make this inference, he wants to retrieve all images, all objects, and each image must be viewed as a whole, as an object.

The property of being retrieved as a source of information, or as an object, is not inherent in the images themselves. The same image can be used as a source of data by one user and as an object by another. A panel with calligraphy, for example, can be used as a text of great wisdom by one person, but as a decorative object by another. Nevertheless, most existing image retrieval systems are geared primarily to one kind of retrieval task. Geographic Information Systems (GIS), for instance, are designed for users who are looking for the information the images provide. Yet, a user may also be interested in a map for decorative or sentimental reasons. Similarly, stock photo agencies treat images as objects when they organize their collections. Yet, a user may want, say, a picture of monks meditating in the lotus position to find out how to imitate this position for the purpose of meditation. Whether a retrieval is on the Data or the Objects pole depends entirely on users and the nature of the retrieval task they are carrying out: are they looking for data or for objects?

The Data and Objects poles represent different retrieval tasks and sometimes are likely to bring about opposing characteristics in searching behaviour. A summary of such typical characteristics is given in Table 7.

Relevance criteria. These attributes that users employ to decide whether or not a retrieved image is relevant, but which are not part of the query, are called here *relevance criteria*. A user may, for example, look for a map of Seattle and decide that the first one retrieved is not satisfactory because the printed text is too small to read. Another user may request an image of a rose and find a retrieved photograph highly relevant because the lighting invigorates the colors. The attributes 'the printed text is too small to read,' and 'the lighting invigorates the colors,' are both relevance criteria.

Data Pole	Objects Pole
Images provide information	Images are objects
Relevance criteria can be determined ahead of time	Users will recognize relevance criteria 'when they see them'
Relevance criteria are specifications of which the user is aware	Relevance criteria are latent and are invoked when viewing images
It is possible for users to explain why an image is relevant	It might be difficult for users to explain why an image is relevant
Images can be retrieved with textual and other verbal clues	It might be difficult to find verbal clues for retrieval, clues are often visual
Color, shape and texture can convey information and therefore important for	No evidence exists that color, shape and texture are important for retrieval
Images must include similar information to satisfy the same need	Two very different images may satisfy the same need
Ofness often equals *aboutness*	*Ofness* is likely to be different from *aboutness*
Biographical attributes are not likely to play a role	Biographical attributes are important for relevance assessment
To satisfy requests may require sets of more than one image	Requests are usually satisfied with one image May not require browsing through the
May not require browsing through the whole answer set	equires browsing through the whole answer set
Browsing is time consuming	Browsing can be done rapidly

TABLE 7: *Summary of searching-behaviour characteristics*

Requests for images on the Data Pole are likely to entail situations in which relevance criteria can be determined ahead of time. An experienced user, for instance, may require that a retrieved map have text of a certain size. At this Pole, relevance criteria are likely to be specifications of which the user is aware. It is the opposite case for the Objects Pole, users are more likely to have difficulties expressing relevance criteria ahead of time, but are more likely to recognize these criteria when they 'see' them. It is unlikely that when looking for an image of a rose, a user would express a desire to retrieve a picture in which the lighting invigorates the colors. But when the lighting in a photograph has such an effect, the user can recognize it and employ this effect as a relevance criteria. On the Objects Pole, relevance criteria are likely to be latent specifications which are invoked by viewing individual images.

Further, while it is possible for users requesting images on the Data Pole to explain why retrieved images are relevant, users on the Objects Pole are likely to have difficulty explaining why images are relevant. One such user might select a drawing of a rose out of all rose drawings because she 'liked it best'—without being able to explain why. On the other hand, when asked why they found a particular map of

Seattle useful, users are likely to be able to express various attributes that made the map useful to them.

Retrieval clues. Requests on the Data Pole can be retrieved with textual clues and other types of verbal clues such as shape, texture or color. A physician may be interested in a slide of a foot that shows a curve of a certain shape to examine the effect of such a curve on the foot as a whole, and a tourist may look for red areas on a weather map to find a region with warm temperature for an upcoming vacation. Conversely, on the Objects Pole, because retrieval clues are often visual, it is likely to be difficult to find verbal clues for retrieval. What verbal clues can one employ to retrieve a good drawing of a rose, or a nice picture of Paris?

More specifically for requests on the Data Pole, color, shape and texture, the attributes commonly used in content-based retrieval, can convey information and therefore might be important for retrieval. At this time, however, we have no evidence that these attributes are likely to be used for retrieval on the Objects Pole. In fact, the sample of 100 requests from a stock photo agency we examined earlier indicated that shape, texture and color are not likely to be useful for retrieval on the Objects Pole.

Discrimination among images. All relevant images must have the same or similar information to satisfy a need when retrived on the Data Pole. On the other hand, relevant images retrieved on the Objects Pole may be very different from one another, yet satisfy the same need.

Ofness* and *aboutness. When images are retrieved as sources of information their *ofness* is likely to equal their *aboutness*. A map of Seattle is of Seattle and about it as well, as is the case with a slide of a normal foot. Conversely, *ofness* and *aboutness* are likely to be different from one another on the Objects Pole. An image of a wire can be about the Internet and a drawing of volcanoes spewing with lava and smoke from top and sides can be about anger.

Biographical attributes. For the user on the Data Pole, it usually makes no difference if a retrieved relevant image is part of a series, if it is an enlargement of another image, who created it, and on what kind of paper it is printed — as long as the image provides all the information the user needs and at the level of accuracy the user requires. That is, containers of information usually do not play a role. On the Objects Pole however, they are likely to be important for retrieval. Here, the same considerations might be of central importance to a user who needs a picture of Paris to display in an exhibition about the city.

Size of answer set. Requests on the Data Pole may require more than one image for the user to glean all the needed information. On the other hand, requests on the Objects Pole are usually satisfied with one image. To find out how to get from one point to another, a user may need

more than one map, but if one needs a picture of Paris, one is looking for a single image. A user, of course, may ask for two pictures of Paris. In this case, however, each picture satisfies a different request. For the first one is just a picture of Paris. The second, however, is *another* picture of the city.

The need for browsing. Retrieval on the Data Pole may not require browsing through the whole answer set. Whether the answer set is constructed with an analytical strategy or through navigation and browsing, once users find the information they need, they usually do not need to view additional images or follow new links.

The situation may be completely opposite during retrieval on the Objects Pole. To select a relevant image, users usually want to browse through the whole answer set so they can select the best image. Here, a user who constructs an answer set through navigation and browsing may want to exhaust all links before selecting the image to use. Further, this selected image may have been retrieved in any step of the navigation process. In other words, if a user were to be shown ahead of time the image she would eventually select as best, it is likely she would still need to view the whole set to make this selection. This observation is not surprising because image retrieval on the Objects Pole is guided by visual clues. The only way a user can employ such clues is by viewing the retrieved of images.

Efforts required for browsing. While retrieval on the Objects Pole requires browsing—sometimes through relatively large sets—it is possible to carry out browsing on this Pole in a relatively short time. This observation is substantiated by Enser's report[5] that researchers assume that relevance assessment in image retrieval is done much more rapidly than in text retrieval. A user may see no difficulty in browsing a set of, say, 50 pictures of Paris to select the one he likes most. On the other hand, it is time consuming to browse on the Data Pole. It would be rather taxing on a user to examine 50 maps to find out if they show how to get from one place to another. Luckily, if the third map provides the information, there is no need for a user to browse through the rest.

Discussion

The purpose of this exploration is to begin to examine some of the implications of the retrieval task for the design and evaluation of image databases. These implications, however, should be examined with some caution.

The characteristics of searching behaviour described above, and their manifestations on each Pole of retrieval tasks, are typical of searching behaviour, rather than absolute. For instance, it may happen that a user with a retrieval task on the Objects Pole expresses explicitly

all relevance criteria, whereas one on the Data Pole is unable to do so and claims: 'I'll recognize it when I see it.' While such situations may occur, they are probably not typical. Recognizing that these observations are not absolute should not lower their value for the design and evaluation of databases, however. In reality, both design and evaluation are most often based on typical characteristics, rather than on absolute ones.

Moreover, it is possible that only a few requests in real-life actually have such extreme retrieval tasks, and that most fall in between. This, or any other such statement, cannot be verified as yet because there are no data available to show the distribution of retrieval tasks in real-life requests. The distinction between the poles is nevertheless useful because in-between requests include elements from each Pole. Since elements from one Pole display characteristics that are opposite to those in the other, it is important to recognize all the various elements in each request. This, in turn, makes it possible to identify the characteristics associated with each element.

In addition, the characteristics described above have not been validated by empirical evidence. No study has been carried out to test, for example, whether retrieval tasks on the Objects Pole are typically for one image, or whether retrieval tasks on the Data Pole typically do not require browsing. Nevertheless, it is not premature to examine the implications of these characteristics for design and evaluation. Currently, both design and evaluation of image databases are charging ahead without the guidance of conceptual considerations. For example, much research effort and financial resources are invested in improving content-based retrieval without an awareness of the situations for which such retrieval is useful. Considering the implications of these characteristics may help create an awareness about the applicability of retrieval tests that are conducted in the laboratory to image retrieval as it occurs in real-life. In particular, it may sharpen awareness to the effects of retrieval tasks.

Among the various considerations for the design and evaluation of database images, this discussion will focus on two questions:

- How useful is relevance feedback for image retrieval?
- Does image retrieval require its own performance measurements?

Relevance feedback. In text retrieval systems, a relevance-feedback facility utilizes relevance judgments made by users to improve retrieval. Is it beneficial to use such facilities in image retrieval? As illustrated in Table 7, the attributes of the relevance criteria on the Data Pole are often the opposite of that on the Objects Pole. Retrieving on the Data Pole, users can describe these criteria ahead of time, they can explain why retrieved images are relevant, and they can retrieve with verbal and other clues. Therefore, it might be beneficial to implement

relevance feedback for requests on this Pole. The Objects Pole, on the other hand, presents no such promise. Here, users can recognize relevance criteria only when they see them, they have difficulties in finding verbal clues for retrieval, and they are not always likely to know why they deemed an image relevant. How then can a retrieval system help them find more relevant images?

Further, relevant images retrieved for a request on the Data Pole are likely to have the same information. Therefore, images a user indicates as relevant during relevance-feedback interaction are likely to have common attributes that can then be used to retrieve additional relevant images, or more images similar to the relevant ones. On the Objects Pole the situation is very different. Images that a user deems relevant might be very different from one another. They may have no common attribute that is visible. It is unlikely that an algorithm will be successful in retrieving additional relevant images because it is quite possible that images that are similar to the relevant ones are not relevant at all.

Designers who develop image databases, therefore, may want to consider the nature of the retrieval task the users will be performing before they install a feedback mechanism. Databases for stock photo agencies, for example, are not promising candidates for a relevance-feedback facility, while those for GIS or clinical slide collections may warrant one.

Performance measurement. The most common retrieval measurements for text retrieval are *precision,* which measures what portion of the retrieved set is relevant, and *recall,* which indicates what portion of the relevant documents is actually retrieved. While these measurements have many shortcomings, they are the only standard measurements for the performance of text-retrieval systems. Moreover, precision and recall have been used in image retrieval tests. Recently, the idea that these measurements might not be adequate for image retrieval has started to spread, but attempts to develop new measurements are limited.

Precision. What is the value of precision to image retrieval? Indirectly, precision reflects the time and effort a user has to invest in order to retrieve the needed information: the higher the precision, the less time and effort are required. Further, there is a basic assumption that users are always interested in saving their time and effort.

To save time and effort on the Data Pole, users are looking for the smallest set of images that can provide the needed information. Once they find the information, they probably need to look no further. That is, users are usually not interested in the set as a whole, but in how long it took them to get to the first image that provided them with the relevant

information. Users on the Data Pole are similar to those whose task is fact retrieval, looking for specific information, such as the population of Seattle, the name of the river crossing Paris, or the average rainfall in Seattle. It seems, therefore, that precision is not a valid measurement on the Data Pole because its calculation is based on the *total* number of relevant items retrieved, and on the *total* number of items retrieved, both of which are of little interest to a user. Instead, evaluators may want to use a measurement that operationalizes the time and effort a user invests to get the first image(s) that provide all the information needed. To that end, evaluators may want to re-examine various measurements, such as Cooper's 'Expected Search Length' measurement[11] and their value for image retrieval (e.g., [12]).

Determining how to save time and effort on the Objects Pole is more complicated. Here, precision has no meaning because there is only one relevant item sought. Precision then does not measure the effectiveness of the retrieval but rather the size of the retrieved set. The smaller the set, the higher the precision. Indeed, the smaller the set that a user must browse through, the less time it takes to find a relevant image or to discover that none is retrieved. The size of the retrieval set, however, is not an adequate measure for the performance of image retrieval, because browsing is required to select the relevant image. Another reason for not accepting the size of the set as a performance measurement is that often users determine a-priori the size of the set, as is the case for some stock photo agencies. In such situations, precision would measure the amount of effort a user is willing to invest in a search, rather than the system's performance.

There could be other ways to measure the system's performance in saving users' time and effort, however. At this point, it seems that instead of evaluating the relevance of individual images, retrieval on the Objects Pole should examine the quality of retrieved *browsing sets* for assessing the relevance of images in the answer set. There are two main reasons for this suggestion.

Because retrieval clues are often non-verbal, and relevance criteria are explicit and latent, it seems almost impossible to index images for retrieval tasks on the Objects Pole. It might be better, therefore, to view indexing of images as a process of assigning membership to a set, where the set is defined by its usefulness for browsing, rather than by subject or another explicit attribute. In other words, indexing is done to create browsing sets rather than for the purpose of differentiating relevant images from non-relevant ones.

Consider, for example, a request for a picture of Paris that a user wants for a travel brochure. In such a case, it makes sense to assume that the user would most benefit from an initial set that includes a large

diversity of pictures of Paris. Fifty images of a bird's-eye view of Paris, or of the Eiffel Tower, will not be useful if the user is not interested particularly in this view or in the Tower. In other words, the most useful set for browsing in this case, might be one that includes images of Paris that are very different from one another.

In addition, to select the relevant image among those retrieved, users often browse through the whole set. It is important, therefore, to retrieve a set that helps users to make this selection with the least amount of effort. Because there is not enough information about human browsing behaviour[13,14], it is difficult to determine what attributes of an image set would decrease the effort in browsing for relevance assessment. Only studies of searching and browsing behaviour that are sensitive to the precision of the *process* of browsing can illuminate the nature of such sets. Indeed, such studies have a great potential to benefit the design and evaluation of image databases.

Recall. What is the value of recall in image retrieval? Based on the assumption that users do not want to miss any relevant item, recall reflects the portion of the relevant items in a database that were not missed in retrieval.

Is it important for users with tasks on the Data Pole to retrieve all relevant images? As explained earlier, users on this Pole are interested in any image(s) that will give them the needed information. The existence of other images with the same information is of no importance. Users will be concerned with recall, however, when no relevant image is retrieved. That is, on the Data Pole the question of recall is scaled down to: can the system retrieve at least one relevant image? It seems, therefore, that with the understanding of image retrieval we have today, the most adequate measurement for image retrieval for tasks on the Data Pole is the effort and time a user invests in finding the first images with the needed information.

The notion of recall may have more meaning for tasks on the Objects Pole. Users might want to see, for instance, all the pictures of Paris before they select one. However, at this time, it is not clear whether browsing through 1,000 pictures of Paris is likely to result in a better selection of a relevant picture than a selection based on 100 pictures. This is because so little is known about image searching and browsing behaviour. Once this behaviour is better understood, it is possible that the notion of recall will be translated, for example, to recall of *browsing sets*, measuring whether no such set of good quality is being missed. Much research and a better understanding of searching behaviour are still needed before the notion of recall can take a valid and meaningful form for retrieval tasks on the Objects Pole.

Conclusions

This initial exploration into image-retrieval tasks reveals that they are likely to have implications for the design and evaluation of image databases. Two extreme tasks — the one a search for information, the other a search for objects—generate different searching behaviours with opposing characteristics. At this early stage, however, there is no empirical evidence on which to base further explorations. It seems that research to validate the existence of the Data and Objects poles, and the characteristics of searching behaviour associated with each retrieval task, are promising avenues for research. Once researchers better understand searching behaviour, they can increase the effectiveness of retrieval systems and develop more useful performance measurements.

While gaining understanding of searching behaviour in image retrieval might be a long process, research in information retrieval can already benefit from the polar construct presented here. Retrieval tests can use it as a conceptual framework for laboratory tests. With such a framework, test requests, as well as assumptions about searching behaviour, can be selected systematically. The more that knowledge about searching behaviour is integrated into laboratory tests, the more their results will be applicable to real-life situations.

Most importantly for such tests, the discussion here indicates that precision and recall as used for text retrieval, might not be adequate tests in image retrieval. Adapting these measurements to image retrieval, or creating new performance measurements, will require both empirical and theoretical explorations. Precision, for example, might be converted on the Data Pole to measure the time and effort a user invests to find the first relevant image, and on the Object Pole to measure the precision of the process of browsing for relevance assessment. Recall, on the other hand, might play a role on both poles only when users cannot find a relevant image. The discussion points to some promising directions. It is hoped that it will also encourage researchers to intensify their efforts to examine performance evaluation for image retrieval.

Acknowledgment

Various people helped me develop the ideas in this paper, most often by challenging them. Special thanks go therefore to my colleague Allyson Carlyle, the participants of the MIRA Dagstuhl Workshop in 1997, and this journal's reviewers.

References

1. CAIVL'97. 1997 *Workshop on Content-Based Access of Image and Video Libraries.* New York, IEEE Press, 1997.
2. JAIN, R. Ed. Visual Information Management. *Communications of the ACM, 40*(12), 1997, 30-80.

3. MAYBURY, M.T. Ed. *Intelligent multimedia information retrieval.* Menlo Park, CA: AAAI Press, 1997.

4. RASMUSSEN, E.M. Indexing multimedia: Images. *Annual Review of Information Science and Technology*, 32, 1997.

5. ENSER, P.G.B. Pictorial information retrieval. *Journal of Documentation, 51*(2), 1995, 126-170.

6. LAYNE, S.S. Some issues in the indexing of images. *Journal of the American Society for Information Science, 45*(8), 1994, 583-588.

7. PANOFSKY, E. *Meaning in the visual arts.* Garden City, NY: Doubleday Anchor Books, 1955.

8. ENSER, P.G.B. Query analysis in a visual information retrieval context. *Journal of Document & Text management, 1*(1), 1993, 25-52.

9. KORF VIDAL, N. Experimental image taxonomy: An inquiry into spontaneous image organization. A thesis for the Degree of Master of Science, 1995.

10. JORGENSEN, C. Image attributes: An investigation. Ann Arbor, MI: UMI Dissertation Services, 1995.

11. COOPER, W. S. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation, 19*(1), 1968, 30-41.

12. DUNLOP, M.D. Time, relevance and interaction modelling for information retrieval. *In:* N.J. Belkin, A.D. Narasimhalu *and* P. Willett, eds. *Proceedings of the 20th Annual International ACM SIGIR* Conference on Research and Development in Information Retrieval, Philadelphia, PA, July 27-31, 1997. New York: ACM Press, 1997, 206-213.

13. CHANG, S.J. *and* RICE, R.E. Browsing: A multidemintional framework. *Annual Review of Information Science and Technology, 28,* 1993, 231-276.

14. KURTH, M. *and* PETERS, T. Browsing in information systems. *Library Hi Tech Bibliography,* 10, 1995.

15. *Dublin Core Metadata Element Set: Reference Description.*
http://purl.org/metadata/dublin_core_elements

16. *Anglo-American Cataloguing Rules,* 2nd ed., 1988 Revision. Eds. M. Gorman & P.W. Winkler. Ottawa: Canadian Library Association, 1988. Chapter 8, Graphic Materials, pp. 200- 219.

17. ARMITAGE, L.H. *and* ENSER, P.G.B. Analysis of user need in image archives. *Journal of Information Science, 23*(4), 1997, 287-99.

Offprint from:
THE NEW REVIEW OF HYPERMEDIA & MULTIMEDIA
Volume 3, 1997
Subscription price for Volume 3, 1997:£70.00/US$130.00
Published annually. Back issues available.

Taylor Graham Publishing, 500 Chesham House, 150 Regent Street, London W1R 5FA, UK.