In IS&T/SPIE Proceedings of SPIE Vol. #6061, Internet Imaging VII, San Jose, CA, USA, Jan 18-1, 2006.

Human Factors in Automatic Image Retrieval System Design and Evaluation

Alejandro Jaimes
FXPAL Japan, Fuji Xerox Co., Ltd.

ABSTRACT

Image retrieval is a human-centered task: images are created by people and are ultimately accessed and used by people for human-related activities. In designing image retrieval systems and algorithms, or measuring their performance, it is therefore imperative to consider the conditions that surround both the indexing of image content and the retrieval. This includes examining the different levels of interpretation for retrieval, possible search strategies, and image uses. Furthermore, we must consider different levels of similarity and the role of human factors such as culture, memory, and personal context. This paper takes a human-centered perspective in outlining levels of description, types of users, search strategies, image uses, and human factors that affect the construction and evaluation of automatic content-based retrieval systems, such as human memory, context, and subjectivity.

Keywords: Image indexing, content-based retrieval, human-factors.

1. INTRODUCTION

The area of multimedia information retrieval has grown tremendously over the last few years. However, in spite of the great number of publications and techniques to automatically index and retrieve multimedia content, the field has not grown in the sense that widespread use applications have failed to take off. Although in recent years most popular search engines have offered image search, and more recently video search, such functionalities are based only on keyword search—indexing is performed only by automatically analyzing the images' metadata (file name, URL, and surrounding text). On one hand, the approach is unstructured in the sense that most images on the web do not have structured metadata to describe their content, and on the other hand, the textual information used to index the images is often inaccurate and incomplete.

In spite of these problems, image retrieval using keywords and automatically indexed metadata has proved effective for some types of searches, particularly when the metadata used accurately describes the content at the desired level. Clearly, the effectiveness of retrieval depends not only on the metadata description, but also on how the user performs the query, his expectations, and other factors.

In this paper, I will discuss the major human factors in image retrieval and point to future research directions to address the human factors that can facilitate or complicate image retrieval tasks. In particular, I will focus on issues such as levels of description, types of users, search strategies, image uses, and human factors that affect the construction and evaluation of automatic content-based retrieval systems, such as human memory, context, and subjectivity.

1.1. Related Work

The authors of [33] report on elements that should be considered in designing and developing an image retrieval system. The authors of [1] discuss evaluation of information retrieval systems, while [2] focuses on the interface. Librarians at Penn State University conducted a 30-month user study [25][26][27] to estimate needs for interdisciplinary image delivery at that university. Results, indicate, among others, the growing importance of personal collections and in

particular, a desire by some users to share those collections as well as to have access to others' personal collections in addition to public collections. Video retrieval is discussed in [3][8], and image retrieval is discussed in [4][6][7][11][16][17][19][20][28]. Issues related to information retrieval evaluation and search strategies on the web and others are discussed in [13][14][15][16][19][21][22][31][32][34], and annotation and personal digital collection management are discussed in [18][29].

2. LEVELS OF DESCRIPTION

One of the biggest difficulties in image retrieval, whether it is manual or automatic, is that images can be indexed at multiple syntactic and semantic levels. Colours, textures, and patterns can be described locally or globally, and semantics can have many different levels of interpretation depending on the particular user. While the meaning of some aspects of an image may be common to many people, it is also true that a particular meaning actually emerges depending not only on the user, but also on the particular collection. In [30], for instance, the authors argue that the meaning of an image emerges from the user's interaction with the collection. The levels of description which are relevant for a particular scenario, then, seem to depend on the collection itself, as well as on the particular query the user is formulating at a given time.

Issues surrounding the emerging meaning of images and levels of description are tightly linked. An important distinction is that of the "*of*" and "*about*" levels [20][9]: an image may be *of* a physical object, but be highly representative, *about* an emotion, or abstract concept. Although any image can have many levels of interpretation, the relevant levels depend on a context which is often given by personal and social factors. For instance, iconic historic images, as well as religious images are often loaded with symbolic significance. Joe Rosenthal's image of U.S. soldiers raising a flag in Mt. Suribachi during World War II, for instance, is used today in many contexts, and is often associated with victory and freedom. The same image might be described using low-level features (black and white, composition, etc.), generic object descriptions (e.g., soldiers), specific object descriptors (e.g., the soldiers' names: Bradley, Hayes, etc.), or abstract descriptors (e.g., victory, endurance, freedom, etc.).

One of the biggest challenges of building automatic image retrieval systems, then, is indexing the images at the right level of description, *and* ensuring that such level matches the user's interest level. While the semantics of an image may change within a collection, it can also change significantly over time: the meaning of Rosenthal's image has certainly changed since it was made on February 3rd, 1945, and without a doubt the image will evoke different emotions in different individuals. The feelings evoked in those in the photograph, in Japanese soldiers who were captured at the time, and in young people now, can be strikingly different. The problem, however, is not limited to historical photographs and occurs with almost any concept and in particular with almost any query. Consider the following examples. A user is performs queries for a "painting", for "blue", for "george bush", and for "white house". As show in Figure 1, the results can be surprising, even for these simple queries. The query for "painting" can be interpreted to be for art paintings (or for the *noun* painting), or it can be interpreted as an action (the *verb* painting). The query in this case is ambiguous: even if the user is looking for "*a* painting", the query does not specify what the painting should be on (e.g., a canvas, a vase, a wall, or a person—note the body painting images). In query for "blue", the user may be searching for images of different shades of blue (e.g., because he wants to paint his house and wants to compare different types of blue), for images that locally contain the colour blue, or for images that evoke the feeling blue. In the third query, the user may be looking for images of George Bush. Instead, he may get pictures about George Bush's policies or satirical cartoons. Clearly, the same query in 1995 would have most likely returned many images of George Bush senior, highlighting that the relevance of the images also changes with time. Finally, a query for "white house" results in a varied array of images. If the user is looking for images of "The White House" in Washington D.C., he gets a good selection. But he also gets images of women in bikinis, of documents, and other items related to the White House, as well as images of white houses.

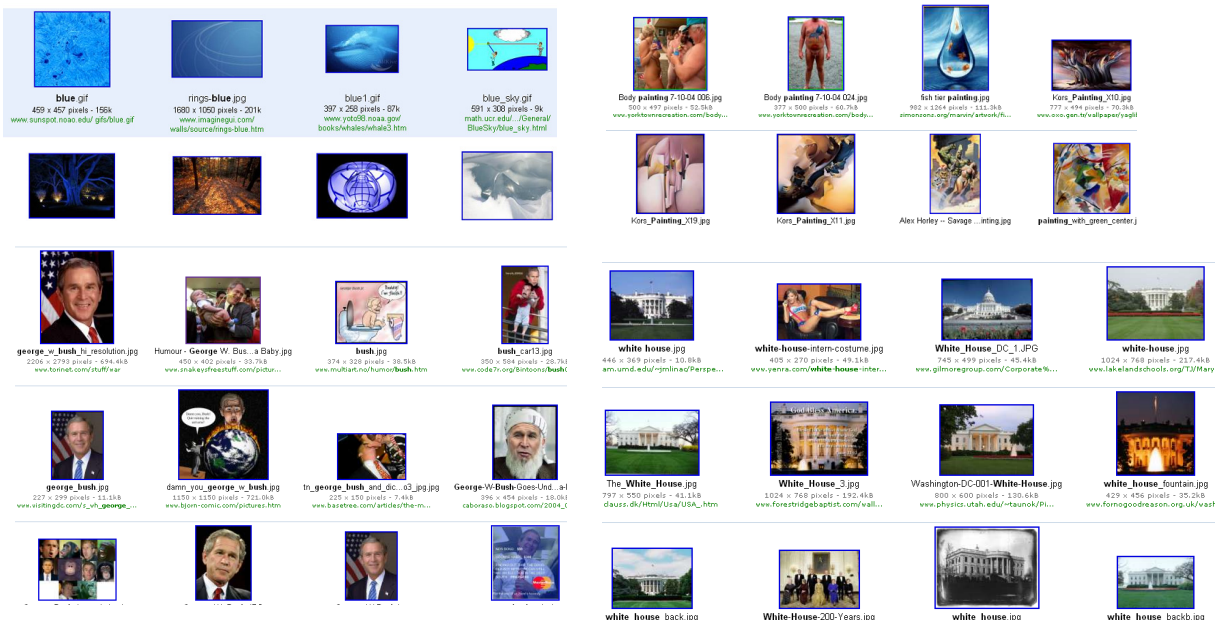

Figure 1. Sample images returned by popular search engines using the queries "blue", "painting", "george bush", and "white house".

The main challenges in terms of levels of description include the following:

- Effectively extracting (and evaluating) features at different levels.
- Obtaining from the user, at query time, an indication of the level of description he refers to.

3. TYPES OF USERS

Users can be classified into different categories depending on the type of search they perform. The query itself depends on several assumptions about the data being searched and the user's knowledge. For instance, let's assume the user is looking for an image X. The user may or may not have seen the image before. If she has seen the image, the search is for a specific item in the database and the problem becomes how to formulate the query to find *that* particular image. If the user has not seen the image, he may be looking for any image of objects, scenes, or events within a category (e.g., cars, people eating, restaurant, etc.). More interestingly, the user may be searching for an image that represents a particular concept, where the concept may again refer to an object, an idealized scene or event, a feeling, or a time period, among many others.

The user, then, can be classified based on whether he has seen the image or not, but more importantly on what he is actually looking for. His *intention* during the search process, however, is different from the actual query. In other words, we must separate what he intends to find and the actual query that he formulates—whether he has seen the image or not, he will formulate a query that depends on factors such as memory and context (described below). Other factors that affect the search strategy include the level of expertise of the user and how familiar he might be with the particular search system: novice, first time visitor, and advanced searcher.

For example, a user may be looking for a particular painting such as Rembrandt's The Night Watch, for Dutch paintings of 17th century, or may be just interested in browsing a particular collection. Clearly, the particular query or search strategy will depend on the user's knowledge and particular task at hand, as well as on the collection. The same user

may utilize different search strategies for different collections. Some researchers ([19][31]), for example, have found that artists and art students often browse an entire collection for information discovery [33]. Advertising companies, on the other hand, often have fairly clear ideas of the concept they wish to represent and look for images that are effective at communicating a particular idea or aspect of a product. Factors such as colour, composition, details in the way people are dressed and the types (and brands) of objects that appear are very important.

In building an image retrieval system, therefore, it is imperative to consider the collection and the particular type of users the system will cater to. In particular, we need to do the following:

- Build systems that adapt to particular users' expertise levels
- Use different evaluation criteria for different types of users (how do we model the user?)

4. TYPES OF SEARCH AND IMAGE USES

In general, there are two types of strategies, one is browsing and the other one is searching. Although many systems provide both functionalities, the two strategies are tightly linked to the particular task at hand and the user often selects only one. Some users will have a very clear and specific idea of what they are looking for, others will have a vaguer idea or concept of what they want. Both types of users will approach the collection in very different ways: the user searching for a specific item may search with a high level of specificity and immediately discard images different from the one he is looking for. The user with a general idea, on the other hand, may formulate a very general query, but spend more time examining results and browsing through the collection to determine if the output satisfies his needs [33]. The type of search chosen will depend on a number of factors including time available, level of expertise, and clarity of information need, among others.

In [33], the following types of user information seeking behaviours are identified:

Prescriptive: used to incorporate prior (e.g. an assignment's) requirements and constraints.
Exploratory: typically used before a specific direction has been developed.
Purposive: more directed and informed searching.
Associative: pro-active search for related and interconnected information to support arguments.
Intuitive: the user is directed by unspecific feelings.
Curious: pursuit of something that piqued interest.
Tangential: clearly beyond prior requirements.
Accidental: accidental actions or system glitches leading to unintended places [34].

The ways users will search for an image is also partially dictated by the intended purpose for which it is to be used, and the following seven broad classes of image use have been identified [1][33]:

Illustration: to represent what is being referred to, e.g. in teaching, with text in a book or journal.
Information processing: the use of the data contained in the image is of primary importance, e.g. in the process of medical diagnosis.
Information dissemination: the image itself contains information that is sought and passed on, e.g. dissemination of a mug shot to police officers.
Learning: knowledge will be gained from the image content, e.g. through research into a topic.
Generation of ideas: images are used to provide inspiration or inspire the creative process.
Aesthetic value: images will be used for purely decorative purposes.
Emotive/Persuasive: an image could be used to stimulate emotions in others or to communicate a particular idea or meaning, e.g. in advertising and media.

Although it is clear that there are differences between the various seeking behaviours and image uses, it is not obvious how these requirements can be implemented in an image retrieval system. The distinction between searching and browsing is very clear, and the system designers can place more emphasis on one or the other. In personal image collections, for instance, it is rare to have annotated images and search by content is difficult because people are often

more interested in finding images of specific people or events (e.g., pictures of uncle Joe at Sam's birthday). This implies that browsing strategies, or systems that make effective use of landmarks (e.g., time structure) can be more effective.

The new wave of community repositories (or images personal image collections made public) also points to new ways of searching and browsing image content. The ability to make comments, tag, and annotate images, leaves the doors open for many indexing and search opportunities combining visual features, textual retrieval, and browsing. Some of the major challenges, however, include the following:

- Facilitate the most suitable searching strategy for a particular problem—adaptable systems?
- Evaluate effectiveness of retrieval systems taking into consideration the search strategy used and the purpose of the search.

Since the majority of image search functionalities are still based on, next I describe some important aspects of search and browsing using text.

4.1. Textual Search and Browsing

Textual search options include open or freetext (all text surrounding the image is used), or keyword-based (only particular words are kept as relevant or important and certain fields are searched). Keyword search can be open (user types query) or guided (user is provided with thesaurus or list). Keywords are often taken from a controlled vocabulary or a list of subject headings, usually referred to as a thesaurus. The particular thesaurus depends on the application and may be a general one such as Art & Architecture Thesaurus (see [11] for a discussion of several thesaurus-based approaches) or a specific one for the collection.

A thesaurus approach has several advantages. It ensures greater consistency in the annotations since different terms with similar or the same meaning are often grouped. It also improves retrieval because the user can select from existing terms in the database and it gives the collection a structure (terms are related to each other by entries that point to narrower, broader, or related terms). For these reasons, thesaurus have been often been the method of choice by librarians for image annotation. One of the problems with this approach, however, is making decisions on what the terms should be and how they are related. In the past, these issues have been dealt with by experts in particular domains—as is the case in the Library of Congress and in specialized collections. More recently, in initiatives such as TRECVID, an ontology is built by a community of experts for automatic annotation. In systems such as Flickr, there is no explicit ontology or thesaurus, but it is clear that the bulk of the manual annotations could be used to "discover" a thesaurus of terms for a particular collection (e.g., find the most frequent terms and their relations).

Although a thesaurus is more difficult to implement for text retrieval, it is inevitable when automatic analysis algorithms are constructed, as typically the types of object or scene classifiers that are implemented are decided in advance. The thesaurus in this case may actually correspond to an ontology, which in its simplest form may correspond to just a list of concepts and associated image representations [9].

Browsing can of course be done using a thesaurus of subject headings or be done based on content. Some of the key challenges include the following:

- Automatically or semi-automatically build user-relevant concept hierarchies and thesauri.
- Evaluate browsing systems using quantitative measures and exploit community annotations.

5. PERSONAL FACTORS

5.1. Memory

Traditional retrieval paradigms assume that the user remembers exactly what she is looking for. If the user is looking for a specific image he has seen previously, it is clear that memory will play a particularly important role in several cases:

(1) the database contains very similar images; (2) the query is performed using query-by-sketch paradigms, and (3) the user has searched for and found the image before and tries to replicate the query.

When the database contains similar images the user may not remember enough details to be able to find the specific image he is searching for. This type of problem is very common particularly in video or image search within very specific collections. For example, in smart conference room applications, the results of video search are often presented in the form of still images. The user may have attended meetings that were recorded and is looking for a particular meeting he attended. Many of the images will be visually similar—the user is unlikely to remember enough details to differentiate from the image alone which video he is looking for. In image retrieval the same problem may arise simply because the user does not remember enough details of the image. For example, a client of a stock image database company browses through a couple of hundred images at a given time while looking for a particular image. Several days, weeks, or months later he needs a new image for a particular project, and remembers that one of the images he saw the first time would be perfect for the project. The user must then return to the database and try to find this image again.

One of the biggest challenges in this case is that the user's memory of the image and the actual image might differ greatly, to the point where the user may overlook the image in question during a subsequent search of the collection.

One of the problems with the current search paradigms is that they completely ignore the possibility of the user not knowing exactly what he is looking for. One way to deal with this issue is to focus the query process on the user rather than on the query itself [9].

5.2. Context and Subjectivity

The semantics of an image depends completely on the context in which it is *viewed*—the semantics of the image is given by the user herself, in what has been referred to as emergent semantics [30]. This interpretation depends on many factors including culture, time, and purpose among others. Cultural factors play a key role in the interpretation of the image at every level. Consider, for instance, the importance of certain colors, patterns, and gestures in different cultures. One of the problems is that these cultural differences are often difficult to quantify and therefore to index.

In performing a query, however, one of the issues is that users have a particular context in mind during the search process. This was illustrated in Figure 1, which shows the results of several simple keyword searches. The user may type a particular word, but the meaning of the word will have a different significance depending on the context. The problem then, is not just that there are different levels of interpretation, but that the context of the search is lost during the query. If the system knew, for example, that the query of "george bush" is for a satirical article in a magazine, then it can return the correct images (if the image index contains that information). A more viable alternative, of course, is to let the user express the context. In practice, this is done by expert users as they reformulate the query after the initial results are obtained, or in relevance-feedback systems that allow the user to better specify his query.

In the examples of Figure 1, we see that there is a big difference between the actual query and the intention of the user. In particular, the meaning of the query is given by the context, which is often not expressed in the query itself but is a part of the mental model the user has in mind when searching. Of course, a great deal of contextual information can also be obtained at the time of capture (e.g., [5]).

Another issue is that the users' interests evolve during information exploration as they learn and discover more about the topic at hand. This can happen either because the user learns about the collection and modifies his query, or because he subjectively makes changes to his query as he searches. For instance, the user is looking for an image of george bush, but as he sees the results he finds unexpected images and changes his mind on the type of image that he wants to use. In this case the search becomes a way of browsing the collection, and subjectivity makes him change his navigation strategy. Subjectivity, of course, can be very high amongst different individuals, but more importantly, can also be an important factor for the same individual at different times. As discussed above, the meaning of images changes over time, not because the images themselves change, but because new information about the images or events that they depict can influence the feelings they evoke or their significance. Therefore, there is a tight link between subjectivity

and context, and in general we can say that context is broader and applies to a group, where as subjectivity tends to be more arbitrary and depend on individuals.

The key issues in system construction and evaluation in terms of context and subjectivity include the following:

- Build new methods that leverage human factors such as memory and subjectivity.
- Model cultural and social factors, as well as the changing nature of image descriptions depending on context

6. CONCLUSIONS AND FUTURE WORK

Research in content-based image retrieval has been an exciting field of growing research in the last few years. In spite of this, the field remains in its infancy in the sense that automatic content-based analysis for image retrieval is not in widespread use as most of the commercially available systems (for web search and others) rely only on text. One of the possible reasons for this might be that most of the work in content-based analysis has focused mainly on low-level features or on the detection of specific concepts (e.g., face, car, indoor, outdoor, etc.), largely ignoring the final user of the system, and more importantly, the human factors that pertain to image indexing, browsing and retrieval.

In this paper I have given an overview of some of the factors to consider when building and evaluating an automatic image retrieval system. In particular, the discussion has focused on levels of description, types of users, search strategies, and image uses. In addition, I have discussed human factors that affect the construction and evaluation of automatic content-based retrieval systems, namely human memory, context, and subjectivity.

There is much work to be done in this area, and evaluation remains an important issue. On one hand, we must continue improving our algorithms to automatically detect concepts by evaluating them on common (public) image database collections and common retrieval tasks. On the other hand, we must work from a human-centered approach and have a good understanding of the human factors that affect they way images are searched for and browsed. Clearly, search strategies depend on particular collections and on particular user needs, context, and limitations.

In order to succeed in building image retrieval systems that use automatic content-based analysis, therefore, we must take a holistic approach and find ways to model all of the relevant variables pertaining to the final, *human* users of theses systems. For this reason, the discussion in the paper have not been limited to visual features—they are just one part of the spectrum in the image retrieval problem.

7. REFERENCES

[1] N. Belkin, S. Dumais, J. Scholtz and R. Wilkinson, “Evaluating interactive information retrieval systems: Opportunities and challenges,” In Proceedings of *CHI 2004*, pp. 1594-1595, 2004

[2] G. Brajnik, S. Mizzaro, and C. Tasso, “Evaluating user interfaces to information retrieval systems: A case study on user support,” in proc. of *ACM SIGIR1996*.

[3] M.G. Christel and A.G. Hauptmann, “The Use and Utility of High-Level Semantic Features in Video Retrieval,” in proc. *CIVR 2005,* Singapore.

[4] L.R. Conniss, A.J. Ashford, and M.E. Graham, “Information seeking behaviour in image retrieval: VISOR I final report,” (Library and Information Commission Research Report: 95). *Newcastle: Institute for Image Data Research*, 2000.

[5] M. Davis, M. Smith, J. Canny, N. Good, S. King, and R. Janakiraman. "Towards Context-Aware Face Recognition." In *proceedings of 13th Annual ACM International Conference on Multimedia (MM 2005)* in Singapore, ACM Press, 483-486, 2005.

[6] R. Fidel, “The image retrieval task: Implications for design and evaluation of image databases,” *New Review of Hypermedia and Multimedia*, Vol. 3. 181-199, 1997.

[7] L. Hollink, A.Th. Schreiber, B. Wielinga, M. Worring, "Classification of User Image Descriptions," *International Journal of Human Computer Studies* 61/5, pp. 601-626, 2004.
[8] L. Hollink, G.P. Nguyen, D. Koelma, A.Th. Schreiber, and M. Worring, "User Strategies in Video Retrieval: A Case Study," *CIVR 2004*: 6-14, 2004.
[9] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata, "Memory Cues for Meeting Video Retrieval," *1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences in conjunction with ACM Multimedia 2004*, New York, NY, USA, October 2004.
[10] A. Jaimes and J.R. Smith, "Semi-Automatic, Data-Driven Construction of Multimedia Ontologies," *ICME 2003*, Baltimore, USA, 2003.
[11] A. Jaimes. *Conceptual Structures and Computational Methods for Indexing and Organization of Visual Information*, Ph.D. Thesis, Department of Electrical Engineering, Columbia University, February 2003.
[12] A. Jaimes and S.-F. Chang, "A Conceptual Framework for Indexing Visual Information at Multiple Levels", in *Internet Imaging 2000, IS&T/SPIE*, San Jose, CA, January 2000.
[13] B.J. Jansen, A. Spink, J. Bateman, & T. Saracevic, "Real life information retrieval: A study of user queries on the web," *SIGIR Forum*, Vol. 32. No. 1., pp. 5 –17, 1998.
[14] B.J. Jansen, and U. Pooch, "Web user studies: A review and framework for future work," *Journal of the American Society of Information Science and Technology*, 52(3), 235 – 246, 2000.
[15] U. Lee, and Z. Liu "Automatic Identification of User Goals in Web Search," in *proceedings of WWW 2005*, Tokyo, Japan, 2005.
[16] S. McDonald, and J. Tait, "Search strategies in content-based image retrieval," in *SIGIR 2003*, Toronto, Canada, July 28-Aug. 1, 2003.
[17] C. Jörgensen and P. Jörgensen, "Image querying by image professionals: Research Articles," *Journal of the American Society for Information Science and Technology*, Vol. 56, Issue 12, Pages: 1346 – 1359, October 2005.
[18] Kustanowitz, J. and Shneiderman, B., "Motivating Annotation for Personal Digital Photo Libraries: Lowering Barriers While Raising Incentives", *Univ. of Maryland Technical Report HCIL-2004-18*, January 2005.
[19] S.S. Layne, "Artists, art historians and visual art information," *Reference Librarian*, 47, 23-36, 1994.
[20] S. S. Layne, "Some Issues in the Indexing of Images," *JASIS* 45(8): 583-588, 1994.
[21] G. Marchionini, G. "Interfaces for end-user information seeking," *Journal of the American Society for Information Science*, 43(2), 156-163, 1992.
[22] G. Marchionini, "Information seeking in electronic environments," Cambridge: University Press of Cambridge, 1995.
[23] National Portrait Gallery (http://www.npg.org.uk/live/search), accessed Nov. 4, 2005.
[24] R. Navarro-Prieto, M. Scaife, and Y. Rogers, "Cognitive Strategies in Web Searching," *5th Conference on Human Factors & the Web*, 1999.
[25] Penn State University Libraries Visual Image User Study (VIUS) (http://www.libraries.psu.edu/vius/), accessed Nov. 4, 2005.
[26] H. Pisciotta, R. Brisson, E. Ferrin, M. Dooris, and A. Spink, "Penn State Visual Image User Study," *D-Lib Magazine*, Vol. 7, No. 7/8, July/August 2001.
[27] H.A. Pisciotta "Penn State's Visual Image User Study" in *Portal: Libraries and the Academy,* Vol. 5, No. 1, pp. 33-58, January 2005.
[28] E. Rasmussen, "Indexing images," *Annual Review of Information Science and Technology*, Vol. 32. Medford, NJ: Information Today. 169-196, 1997.
[29] K. Rodden, and K.R. Wood, "How Do People Manage Their Digital Photographs?", in the *ACM Conference on Human Factors in Computing Systems (ACM CHI 2003)*, Fort Lauderdale, Florida, April 2003.

[30] S. Santini, A. Gupta, and R. Jain, "Emergent Semantics through Interaction in Image Databases," *IEEE Trans. Knowl. Data Eng*. 13(3): 337-351, 2001.
[31] D. Stam. "Artists and art libraries," *Art Libraries Journal*, 20(2), 21-24, 1995.
[32] D.E. Rose and D. Levinson, "Understanding user goals in web search", *WWW 2004*, New York City, May 17-22, 2004.
[33] Technical Advisory Service for Images (http://www.tasi.ac.uk/), accessed November 5, 2005.
[34] S.C. Yang, "Qualitative exploration of learners' information-seeking processes using Perseus hypermedia system," *Journal of the American Society for Information Science*, 48(7), 667-669, 1997.