Faceted Metadata for Image Search and Browsing

Ka-Ping Yee, Kirsten Swearingen, Kevin Li, Marti Hearst
CS and SIMS, UC Berkeley
Berkeley, CA 94720
{ping,kirstens,kevinl,hearst}@sims.berkeley.edu

ABSTRACT

There are currently two dominant interface types for searching and browsing large image collections: keyword-based search, and searching by overall similarity to sample images. This paper presents an alternative in which users are able to navigate explicitly along conceptual dimensions that describe the images. The interface makes use of hierarchical faceted metadata and dynamically generated query previews. A usability study, conducted with 32 art history students exploring a collection of 35,000 fine arts images, compares this approach to a standard image search interface. Despite the unfamiliarity and power of the interface (attributes which often lead to rejection of new search interfaces), the study results show that 90% of the participants preferred the metadata approach overall, 97% said that it helped them learn more about the collection, 75% found it more flexible and 72% found it easier to use than a standard baseline system. These results indicate that that a category-based approach is a successful way to provide access to image collections.

KEYWORDS: Search User Interfaces, Faceted Metadata, Image Search

INTRODUCTION

Image collections are rapidly coming online, and many researchers have developed user interfaces for browsing and searching those collections. Probably the most familiar image search interface today is that used by web image search engines in which users enter keyword terms and images are shown in a table, ordered by some measure of relevance. These systems can be effective for searching for very specific items, but do not support browsing and exploratory tasks well [10, 7, 9]. By contrast, many research systems approach image retrieval by analyzing images in terms of their visual properties such as which colors and textures are present. However, results of usability studies call into question the usefulness of searching for images according to low-level visual properties [15, 10].

By contrast, and perhaps counter-intuitively, ethnographic studies indicate that professionals who look for images on a regular basis (i.e., journalists, designers, and art directors) want to browse and search images using textual category labels [5, 10, 7, 1]. Despite this, few image search engines provide the ability to navigate images by rich category sets, and those that do often make use of unwieldy interfaces [10].

In this paper we present an interface for large image collections which allows users to navigate explicitly along conceptual dimensions that describe the images. The interface makes use of hierarchical faceted metadata (described below) and dynamically generated query previews [14], to seamlessly integrate category browsing with keyword searching. To arrive at the current design, our team conducted several rounds of usability studies and interface redesign [8]. This paper presents the results of a new usability study whose goal is to directly compare the faceted category design to the current most popular approach to image search. Conducted with 32 art history students using a fine arts image collection, the study found strong preference results for the faceted category interface over that of the baseline, suggesting this is a promising direction for image search interfaces.

The remainder of the paper describes related work, the faceted metadata, the category-based interface design, the baseline interface, and the study design and results, concluding with a discussion of the larger lessons that can be drawn from this effort.

RELATED WORK

The bulk of image retrieval research falls under the rubric of "content-based" image retrieval; the term refers to systems that perform image analysis in order to extract low-level visual properties of the images. These include color and texture analysis [12, 13] and object segmentation [4]. Some systems also incorporate information extracted from associated text [17]. A good summary of content-based image retrieval can be found in [18].

There has been a great deal of research on these systems, but only a small subset of this work has included usability studies. Rodden et al. [15] performed a series of experiments whose goal was to determine if and how organization by visual similarity is useful, using as features global image properties (colors and textures) and the spatial layout of image regions. Their results suggested that images organized by cat-

egory labels were more understandable than those grouped by visual features.

Ethnographic studies of image search needs indicate that there is a great need for more conceptually rich image search. In a study of art directors, art buyers, and stock photo researchers [7], Garber & Grunes found that the search for appropriate images is an iterative process; after specifying and weighting criteria, searchers view retrieved images, and then add criteria, add restrictions, change criteria, or redefine the search. The concept often starts out loosely defined and becomes more refined as the process continues.

Markkula and Sormunen [10] report on a field study of journalists and newspaper editors choosing photos from a digital archive, for the purposes of illustrating newspaper articles. Journalists stressed the need for browsing, and considered searching for photos of specific objects to be a "trivial task." Selection of search keys for general topics was considered difficult; journalists emphasized the need for photos dealing with places, types of objects, and themes. The journalists had access to an "advanced search" interface that allowed them to search on many different features at once, but its format, which consisted of about 40 entry forms and drop-down boxes, was seen as too complex, and was rarely used. Thus, although they had the desire to do searches on multiple categories, the interface discouraged them from doing so.

A query study also supports the notion that users want to search images according to combinations of topical categories. Armitage and Enser [1] analyzed a set of 1,749 queries submitted to 7 image and film libraries. They classified the queries into a 3 x 4 facet matrix; for example, *Rio Carnivals* falls under *Geographic Location* and *Kind of Event.* They do not summarize how many queries contain multiple facets, but show a set of 45 selected queries, to which they have assigned an average of 1.9 facets per query.

The system proposed by Garber & Grunes [7] is the interface most similar to our approach. The interface operated in two modes (i) by showing the metadata associated with a selected image, and presenting images in an order that reflects the number of categories they have in common with the target image, and (ii) allowing the user to select a set of category labels, and showing sample images for similar categories (e.g., showing images labeled *New England, Africa*, and *Egypt* when the category label *Florida* is selected). Hierarchy information was not shown, and no information was provided about how many images are available in each category. Focus groups observing the demonstration were very enthusiastic about it, but no followup work appears to have been done.

METADATA

Below we define and illustrate the notion of faceted metadata.

Faceted Metadata

Content-oriented category metadata has become more prevalent in the last few years. Many individual collections already have rich metadata assigned to their contents; for example, biomedical journal articles have on average a dozen or more content attributes attached to them. Metadata for organizing content collections can be classified along several dimensions:

- The metadata may be *faceted*, that is, composed of orthogonal sets of categories. For example, in the domain of fine arts images, facets could include Media (etching, woodblock, ceramic, etc.) Themes (Military, Religion, The Arts, etc.), Places (Bridges, Buildings, Roads, etc.) Materials, Locations, Periods, and so on.
- The metadata (or an individual facet) may be *hierarchical* ("located in Vienna, Austria, Europe") or *flat* ("by Pablo Picasso").
- The metadata (or an individual facet) may be *single-valued* or *multi-valued*. That is, the data may be constrained so that at most one value can be assigned to an item ("measures 36 cm tall") or it may allow multiple values to be assigned to an item ("uses oil paint, ink, and watercolor").

There are a number of issues associated with metadata, including which descriptors are correct or at least most appropriate for a collection of information, and how to assign metadata descriptors to items that currently do not have metadata assigned. Many researchers are investigating this (e.g., [17]), and there are in fact many existing, important collections whose contents have hierarchical metadata already assigned.

Collection Preparation

The collection under study contains the approximately 35,000 images out of the more than 82,000 that are part of the Thinker collection of the Fine Arts Museum of San Francisco (metadata was available only for a subset of images). This collection contained standard arts metadata facets, including artist name, type of media (etching, aquatint, etc.), and date. However, there was little in the way of content-based metadata, that is, no metadata categories that described the appearance of items or the images depicted in them, as in the case of paintings. Many of the images did, however, have sentential or phrasal descriptions of their contents. For example:

- *A man riding in cart drawn by two horses.*
- *soup can, not in traditional colors: i.e. green lid, purple and orange lettering, etc.; Campbell's condensed tomato soup in purple, aqua and orange on purple background.*

We developed an algorithm to semi-automatically convert these descriptions into a set of metadata categories that were assumed to be useful for students and scholars of art history.

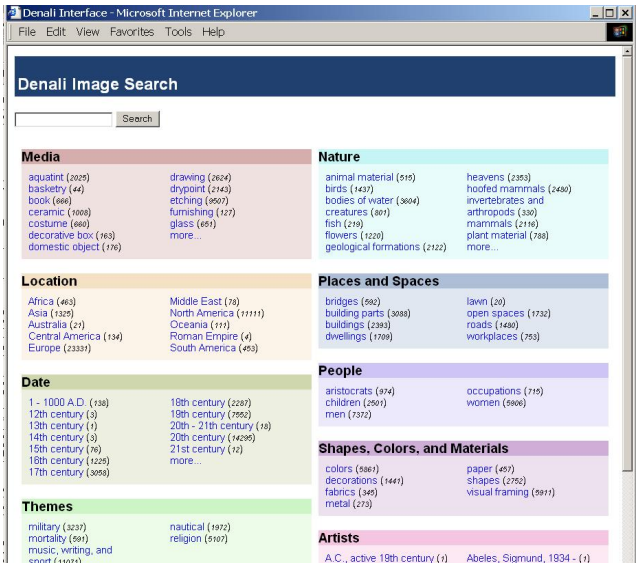

Figure 1: The opening page shows a text search box and the first level of metadata terms. Hovering over a facet name yields a tooltip (here shown below Locations) explaining the meaning of the facet.

This was done by comparing the words in the descriptions to the higher-level category labels in WordNet [6], and retaining a subset of the most frequently occuring categories. Certain categories tend to correspond to highly ambiguous terms (e.g., “arm”, “head” and other body part terms) and so were discarded. Many other ambiguous words (such as “punt”) only have one sense in the collection and so could be retained. Although some labels were incorrectly assigned, the algorithm worked surprisingly well. Usability study participants occasionally commented on the incongruities between the label and the image, but they still appeared to trust and like the category labels. Many participants expressed pleasure at seeing content descriptors in addition to the traditional descriptors of who, what, and where. The leaf-level category labels were manually organized into a set of hierarchical facets, using breadth and depth guidelines similar to those found in [2].

INTERFACE DESIGN

The Faceted Category Interface

Unifying Goals

Our design goals are to support search usability guidelines [16], while avoiding negative consequences like empty result sets or feelings of being lost. Because searching and browsing are useful for different types of tasks, our design strives to seamlessly integrate both searching and browsing functionality throughout. Results can be selected by keyword search, by pre-assigned metadata terms, or by a combination of both. Throughout the interface, each facet is associated with a particular hue. As a visual cue, categories, query terms, and item groups in each facet are shown in lightly shaded boxes of the appropriate hue. Colors for different parts of the interface are computed by adjusting value and saturation, but maintaining a fixed hue.

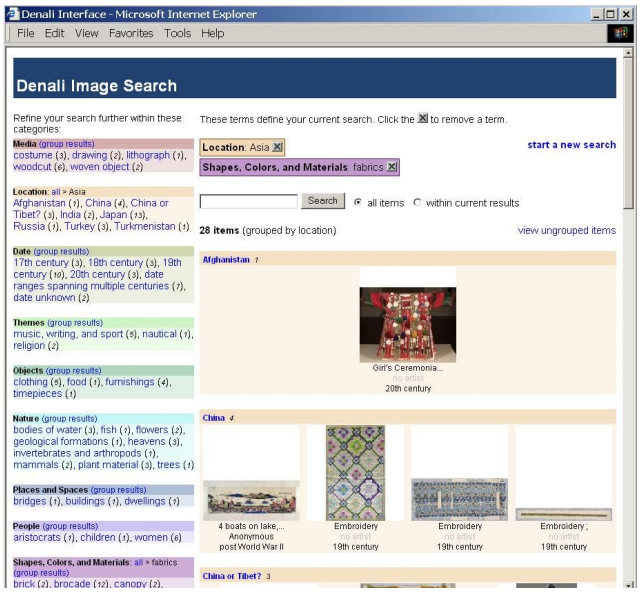

Figure 2: Middle game (items grouped by location).

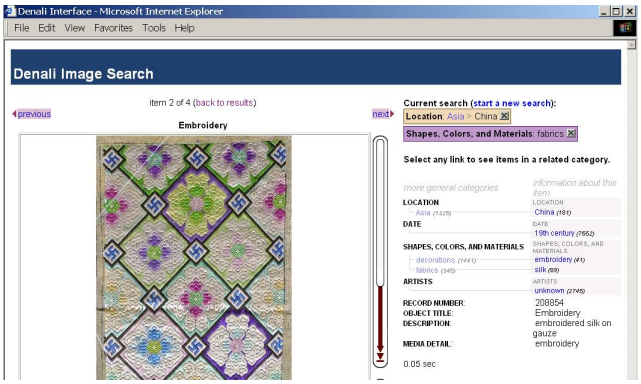

Figure 3: View of an individual item, with contextualized links for expanding the query in several conceptual directions.

In working with a large collection of items and a large number of metadata terms, it is essential to avoid overwhelming the user with complexity. We do this by keeping results organized, by using simple point-and-click interactions instead of imposing any special query syntax on the user, and by not showing any links that would lead to zero results. Every hyperlink that selects a new result set is displayed with a query preview (an indicator of the number of results to expect).

The design can be thought of as having three phases, by loose analogy to a game of chess: the opening, the middle game, and the endgame. Although the most natural progression is to proceed through these phases in order, users are not forced to do so.

Opening

The primary aims of the opening phase are to present a broad overview of the entire collection and to allow many starting paths for exploration. The opening page (Figure 1) displays each metadata facet along with its top-level categories. This provides many navigation possibilities, while immediately

familiarizing the user with the high-level information structure of the collection. The opening also provides a text box for entering keyword searches, giving the user the freedom to choose between starting by searching or browsing.

Selecting a category or entering a keyword gathers an initial result set of matching items for further refinement, and brings the user into the middle game.

Middle Game

The middle game (Figure 2) is the phase during which the result set is evaluated and manipulated, usually to narrow it down. There are three main parts of this display: the result set, which occupies most of the page; the category terms that apply to the items in the result set, which are listed along the left by facet (we refer to this category listing as The Matrix); and the current query, which is shown at the top. A search box remains available (for searching within the current result set or within the entire collection), and a link provides a way to return to the opening.

The key aim here is organization, so the design provides flexible methods of organizing the results. The items in the result set can be sorted on a number of fields, or they can be grouped in categories by any facet. Selecting a category both narrows the result set and organizes the result set in terms of the newly selected facet. For example, suppose a user is currently looking at images resulting from selecting the category *Bridges* from the *Places* facet. If they then select *Europe* from the *Location* facet, not only is the category *Europe* added to the query, but the results are organized by the subcategories of *Europe*, namely *England, Italy*, and so on. Removing or generalizing a category term broadens the result set. Selecting an individual item takes the user to the endgame.

Endgame

The endgame (Figure 3) shows a single selected item, in the context of the current query. Next to the item, the query terms are displayed, together with an innovative hybrid-tree layout that shows all of the metadata terms assigned to the item and their locations within their hierarchies. This layout combines a simple attribute list in the right-hand column, where the actual assigned terms can be quickly read off, with an outline tree view in the left-hand column, where each term is situated in its context within a metadata hierarchy. Selecting a metadata term switches to a new query showing all the items associated with just that term.

This view exposes metadata terms of interest, while also making it easy to navigate *laterally* through the collection. After refining a query in the middle game, a user can head in a totally new direction by selecting an item of interest and then selecting related terms in the endgame.

Keyword Matching

Each item is associated with the text of all its metadata, as well as any additional collection-specific text. The result set formed by a keyword search then contains all items whose text contains the keyword. Keyword search terms can be freely intersected with metadata query terms. In response to a keyword search, an additional panel appears at the top of the middle game display. This disambiguation panel lists all the metadata terms that contain the keyword, with the keyword highlighted in color wherever it appears. The user can select one of these terms to replace the keyword query term with a particular metadata term, or ignore the panel and continue to browse, leaving the keyword term in their query.

System	Collection	Results Per Page	Show Cats?	# Used
Google	Images from web	20	No	27
AltaVista	Images from web	15	No	8
Corbis	Photos	9-36	No	8
Getty	Photos, art	12-90	Yes	6
MS Office	Clip art, photos	6-100	Yes	NA
Thinker	Fine arts images	10	Yes	4
Baseline	Fine arts images	40	Yes	NA

Table 1: Comparison of features in popular existing image search interfaces. Show Cats indicates the display of hyperlinked categories when images are selected.

Intermediate Listings

When a particular query yields too many items or too many subcategories to display at once, an intermediate page is shown, listing all the subcategories and suggesting that the user choose one. The subcategories are listed in columns and grouped in alphabetical order.

The system is built using Python, MySQL, and the WebWare toolkit[1]. All components of the interface are dynamically generated, based on the facets and facet values defined in a relational database. Query previews are generated using the SQL COUNT(*)/GROUP-BY operator to count the number of items that fall into each subcategory.

The Baseline Interface

Today many users are familiar with keyword-based image search, as embodied by web image search engines. Table 1 compares some of the features of 5 image search engines: Google Image Search, AltaVista Image Search, Corbis, GettyImages, and MS Office Clipart, in addition to The Thinker, the search engine currently available for the art history collection used in the study below.

When the user selects an image for detailed viewing, three systems (GettyImages, MSO, and The Thinker) show related topical category labels, hyperlinked to act as queries (e.g., showing the categories *Flowers, Nature* next to an image of poppies). These categories are not explicitly faceted nor hierarchical, and are usually not shown in any meaningful order.

To create a fair comparison of the two interfaces, we built an image search system that is representative of the best aspects of the five popular image search engines of Table 1. When in doubt we usually opted to make the baseline act as Google image search does, due to its familiarity to the user

[1] www.python.org,www.mysql.com, webware.sourceforge.net

population.

The opening, or start page, for the baseline interface provides an entry form for typing in search terms, an illustrative image, and a two sentence description of the collection (that mimics that of The Thinker) and some information on how to search the collection. We chose to provide a search method in which terms are implicitly ANDed, since this practice has become widely-adopted due to Google's use of it. Only one participant (in the pre-test) asked about doing advanced Boolean queries. Adjacent words enclosed in quotation marks are treated as phrases. Stemming is not used, both because of the confusion it can cause [11], and because Google does not use it.

After the user enters search terms, a linked list of pages of search results is shown, along with a description of how many images were found as a result of the query. The images are shown in a table of 10 rows of 4 images each, in alphabetical order according to image title[2]. The user can click through a page at a time, can do a new query at the search form which appears at the top of the screen (the default is to search the entire collection) or can click on an image to see more detail.

In the detailed view, a larger version of the image is shown along with a listing of its associated metadata. In addition, the baseline has a feature that makes it more powerful than the other keyword search systems. It shows a hyperlinked list of category labels that translate into queries on the corresponding category label in the faceted category interface. For example, if an image has been assigned the category label *Bridge* in the faceted category interface, a hyperlink to a query is also shown in the baseline interface; that hyperlink retrieves all items in the *Bridge* category is shown next to the image. The categories are shown in alphabetical order, but no preview is shown of the number of items in the category. Thus here the baseline interface departs from the Google approach and is similar to the category views provided by other systems of Table 1.

However, because the baseline interface does not need to compute query previews, it is much faster than the faceted category interface. For the studies described below, we measured the average processing time for the category interface to be an order of magnitude longer than that of the baseline interface.

Prior Work

To develop the target interface, we followed standard interface design practice. Beginning with the domain of architectural design, we did an ethnographic study of how architects search for and use images as inspiration for design. This was followed by a cycle of low-fidelity prototyping, informal usability testing, and redesign. After this we conducted two rounds of development and two usability studies. These studies were useful for answering questions about various design features, and whether users would respond well to navigation of multiple simultaneous hierarchical facets. However, up to this point we had not compared the design to a more standard baseline, to determine if this richer method of search would be preferred and more effective over a more standard interface. Thus this paper presents the results of a new study to answer the question: is this design better than the current state of the art in image search interfaces?

[2]It is difficult to determine the ranking algorithm used by the web search engines; presumably it is a function of the match of the query terms to the words surrounding the images in the hypertext. The other systems do not seem to have a ranking function; three systems allow grouping according to broad categorical features such as color vs. black-and-white or media type.

USABILITY STUDY

Participants

Working with participants who are interested in the collection in question has been found to be especially important in search usability studies [3]; this has been our experience as well. We chose to use a fine arts collection for this study because it is possible to recruit art history students and people who have recently taken art courses, as the study participants. Data from 32 participants was used in the analysis. (A pre-test was conducted on three participants and data for two outliers was discarded.) The participants were all regular users of the Internet, searching for information either every day or a few times a week. They searched for images online less frequently, with the majority searching for images less than once per week. Table 1 summarizes their familiarity with various image search systems; 4 people had used the fine arts image collection with its official interface.

Apparatus

Participants received a $15 gift certificate for participating in a session that lasted about 1.5 hours. All participants were tested in a lab setting, using the Internet Explorer v6 browser on Windows 2000 workstations with 21 inch monitors set at 1280 X 1024 pixels in 24-bit color. Data was recorded with multiple methods: (a) server logs (b) behavioral logs (time-stamped observations) and (c) paper surveys after each task, each interface, and at the end of the session. One to two experienced usability analysts conducted the sessions; when two were available, one analyst took written notes while the other facilitated the session. Data from all the sources was collated to create a complete record of the test session.

Design and procedure

The study used a within-subjects design. Each participant used both the faceted category interface (henceforth FC) and the baseline interface; each interface was the starting view for half the participants. The interfaces were assigned neutral names (Denali and Shasta).

In earlier studies we walked the participants through the features of the experimental interfaces. By contrast, and to better mimic the situation that occurs in practice, in this study, participants were not introduced to the features nor told anything in advance about the systems other than that they both accessed the same collection of 35,000 Fine Arts images. We

also informed participants that keyword searching was available in both interfaces and briefly explained the commands they might use to search (an asterisk for wildcard searching and quotation marks for phrases).

Throughout the study, subjective ratings were reported on a 9-point Likert scale, with 1 signaling "strongly disagree," 9 meaning "strongly agree," and 5 meaning neutral. Because we have found participants tend to be generally positive about the current interface, we adopted a wide range in order to have a more sensitive testing instrument.

Tasks

The tasks were designed to reflect the contents of the collection and the art history background of the students. Participants completed four tasks on each interface, two structured and two unstructured:

1. (3 min, unstructured). Search for images of interest.
2. (11-14 min, structured). Gather materials for an art history essay on a given topic. Complete 4 subtasks, ranging from very specific to more open ended e.g., (i) Find all woodcuts created in the United States, (ii) Choose, the decade for which the collection seems to have the most images of US woodcuts, (iii) Select one of the artists who worked during this period and show all of his or her woodcuts, (iv) Choose one of the subjects depicted in these works and find another US woodcut artist who has treated the same subject in a different way.
3. (10 min, structured). Compare related images in order to write an essay (e.g., find images by artists from two different countries that depict conflict between human beings).
4. (5 min, unstructured). Search for images of interest.

Task 2 used metadata categories clearly visible in the start page and matrix of FC. However, we carefully framed the wording of Task 3 so as not to reflect a particular facet. Each of Tasks 2 and 3 had two versions; study design was balanced in terms of which queries were assigned to each interface. At the end of the session, we asked participants whether they felt the structured queries were equally difficult; 30 out of 32 stated that they were equivalent. As a double-check, we looked at the difficulty ratings in the post-task questionnaires for the different tasks; we found no significant differences between the two task sets (both t's < 1.7, both p's > 0.05).

Results

It is difficult to evaluate browsing tasks, since there are no correct answers and since the goal is not necessarily to minimize time used. Thus the tasks and measures were designed to test the following hypotheses about FC:

(i) Participants will experience greater search satisfaction and success in FC than in the Baseline, feel greater confidence in the results, produce higher levels of recall, and encounter fewer dead ends.
(ii) Overall, FC will be perceived to be more useful and flexible than the Baseline.
(iii) Using FC, participants will feel more familiar with the contents of a collection.
(iv) Participants will use FC to create multiple-facet queries during their self-directed searches.

Task Satisfaction and Success

After each structured task, participants completed a short questionnaire. Using FC, participants felt significantly more confident that they had found all of the relevant images in the collection (Task 2: $t(62) = 2.18, p < .05$, Task 3: $t(62) = 2.03, p < .05$) and significantly more satisfied with the results (Task 2: $t(62) = 3.78, p < .001$, Task 3: $t(62) = 2.03, p < .05$) than when they used Baseline (thus supporting hypothesis i).

We evaluated participant success in retrieving all the relevant images for part (a) of Task 2: find all woodcuts created in the United States or all aquatints created in France. In Baseline, 57% of the participants conducting the aquatints task retrieved all the relevant results; in FC, 81% of the participants were successful. For the woodcuts task, 21% of those using Baseline and 77% using FC managed to retrieve all the relevant images (thus supporting hypothesis i). The differences were caused in part by the users of the Baseline not querying both singular and plural forms of words.

Participants indicated they more often found themselves at a dead-end or empty results when using Baseline; this difference was not significant (Task 2: $t(62) = 1.41, p = .163$, Task 3: $t(62) = .499, p = .619$). However, during the structured tasks participants actually did receive empty results in Baseline 82 times while in FC, they received empty results only 26 times (thus supporting hypothesis i).

For search success, we also looked at how many items users opted to bookmark in each system and the usefulness ratings (on a scale from 1 to 10) for those items. In Baseline, participants rated 266 items with an average rating of 8.1; in FC, participants rated 215 items with an average rating of 7.9. In Baseline, participants may have been able to rate more items because the processing speed was so much faster than in FC. The differences in item ratings were not significant ($t(481) = 1.12, p = .26$).

As indicated above, all tasks were assigned time limits but participants were allowed 3 extra minutes on Task 2 when using FC, because of its slower response time. In fact, the query processing time in FC was an order of magnitude slower than Baseline.[3] Participants could complete a task before the time limit had expired. We did not encourage participants to rush through the searches; instead, we asked them to search as they normally would.

Participants spent an average of 9 m, 30 s on Task 2 using

[3]For Task 2, the average processing time per step is 0.3 s for Baseline, but 3.7 for FC. For Task 3, this was 0.37 s for Baseline and 4.3 s for FC.

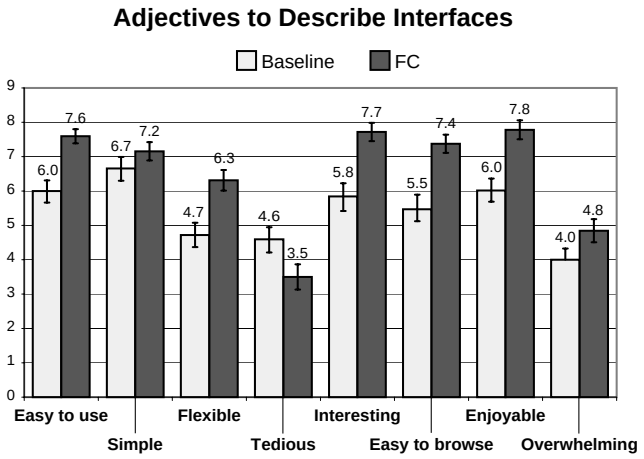

Figure 4: Post-interface assessments; all results statistically significant ($p < .001$) except simple and overwhelming; tedious significant at $p < .05$).

Which interface would you rather use for these tasks?	Baseline	FC
Find images of roses	15	16
Find all works from a certain time period	2	30
Find pictures by 2 artists in the same media	1	29
Overall assessment:	**Baseline**	**FC**
More useful for your usual tasks	4	28
Easiest to use	8	23
Most flexible	6	24
More likely to result in dead-ends	28	3
Helped you learn more	1	31
Overall preference	2	29

Table 2: Post-test preferences for the Faceted Category (FC) Interface vs. the Baseline.

Baseline; in FC, the time spent on this task averaged 12 m, 6 s. For Task 3, participants spent 7 m, 45 s in Baseline and about 9 m in FC. These differences were significant (both p's $< .05$), but may be caused by the slower processing time, and the fact that system errors occurred during 5 of the 32 sessions with FC; restarting the system added time to the tasks. Thus FC did not result in faster usage times; however, we had not hypothesized that it would do so, given that success in browsing tasks is not reflected by faster completion times.

Post-Test Interface Comparison

In the post-interface assessment, stronger differences emerged. Immediately after completing the fourth task on an interface, participants completed an interface evaluation. FC received more positive ratings than Baseline for nearly every measure, as shown in Figure 4. Noteworthy ratings are those for "easy to use" and "easy to browse." Given FC's complex screen design, it is remarkable that users assigned it a rating of 7.6 on average for "simple." Similarly, the fact that FC was not rated to be significantly more "overwhelming" than Baseline [$t(62) = 1.79, p > .05$] testifies to the success of the design. Participants indicated they were more likely to use FC in the future [$t(62) = -3.75, p < .001$]. They also felt more familiar with the collection [$t(62) = -2.17, p < .05$]. These results support hypotheses i and ii.

The order in which interfaces were viewed had a strong effect on these ratings. When FC was viewed first, the interface ratings for Baseline were considerable lower than when Baseline was the first interface shown [$t(26) = 2.67, p < .01$]. The ratings for FC were not significantly affected by being viewed after Baseline [$t(26) = -0.27, p = .783$].

Participants were also asked to compare Baseline to FC and indicate which interface they preferred for different situations (see Table 2). For finding images of roses (a simple, single facet task), about 50% preferred Baseline. However, for every other type of searching, FC was preferred: 88% said that FC was more useful for the types of searching they usually do and 91% said they preferred FC to Baseline overall. Those who preferred the Baseline commented on its simpliciity and stated that the categories felt too restrictive.

Facet Usage

Facet usage in the structured tasks was driven largely by the task content, causing participants to focus on Date, Location, Media, Artist and Theme. However, for the unstructured searches, usage was more evenly distributed across all the facets. Artists (17%), Date (15%) and Location (15%) were the most used facets on the start page, but the 111 starts occurred in the other facets with percentages ranging from 5 - 12%. For refining queries, again Artist (20%), Date (14%), and Location (19%) were most used, but the other facets were used for 6 - 11% of the refining actions (n=139). In the endgame, participants opted to create a new query by clicking on Artist 39%, Media 29%, and Shapes 19% of the time (n=21).

The number of facets used simultaneously was also of interest to us, since this is a unique aspect of FC. In the unstructured tasks, participants constructed queries from multiple facets 19% of the time; in the structured tasks, 45% of the time, thus supporting hypothesis iv. However, when browsing only a single facet, participants frequently used "search within results" to refine their searches (15% for unstructured, 50% for structured).

Qualitative Observations

Users of the Baseline commented favorably on its simplicity and similarity to Google image search, but also noted that the category hyperlinks made it much easier to use.

Many participant reactions to FC followed a pattern. When shown the starting page, more than half explicitly remarked on it, noting that it was "well-organized" and gave them "ideas about what to search for." The query previews were a key ingredient for 9 users, who offered unsolicited comments on this feature's usefulness: "Because of the way it starts, the collection seems more complete because I can tell how many are available in different categories from the front

page."

Once participants tried their first queries, more than half of them commented negatively on the speed. Some wondered aloud about the cause of the slowness, a few said it was "frustrating" and "annoying" and one person commented "at this point, I would go to a different search engine." From the middle game, more than half of the participants explicitly remarked on the matrix, saying favorable things such as it "prompted" them about where to go next. They also generally liked seeing the images grouped into categories: "it does a lot of the work for you, the searching and the categorizing." Three were confused about how the matrix functioned–they thought it was a repetition of the first page and did not realize they could use it to refine their existing query. All other participants did understand the matrix and stated they felt more confident in the results they obtained by browsing. Participants liked having category links in the endgame of both interfaces, but 9 out of 32 explicitly commented on the level of detail in FC, stating that the information here is "useful" and "very clear," "guiding" them through a search.

As participants continued to use the interface, they became more comfortable with it. As an example interaction sequence, one participant began Task 3 (compare images on conflict between peoples) by clicking on *military* at the start page, then refining from an intermediate page to choose *war*. Since there were 824 results, he refined his search further by doing a keyword search within results for *sword*, reducing the number of images to 74. He grouped the results by *artist*, since the task called for him to contrast works by 2 artists. Then he began clicking on images and started formulating his thesis, "This is the Napoleonic view of war–the camera is really far away. Men look like ants and you don't see war itself, the death, just the preparations." It occurred to him that 20th century depictions of war are more graphic. He grouped his 74 results by date and quickly found images by Goya which "zoom in on the misery and suffering" of war.

At the end of the session, participants expressed enthusiasm for the FC interface, wanting to know when the website would be available for them to use. One participant said, "I wish I had this when I was writing papers." They found it "interesting," "enjoyable" and "easy to customize" their searches using the FC interface.

CONCLUSIONS AND FUTURE WORK

We have designed an image access interface that allows users to navigate a large collection using hierarchical faceted metadata in a flexible manner. Despite the fact that the interface was often an order of magnitude slower than a standard baseline, it was strongly preferred by most study participants. These results indicate that a category-based approach is a successful way to provide access to image collections.

We are in the process of developing algorithms to make the query preview generation faster. This is important for future attempts to make the method scale to collections that are 1 or 2 orders of magnitude larger. We also plan in future to perform studies comparing this approach directly to similarity-based approaches, as well as studying the effects of adding personalization, history, and relevance feedback functionality to the design, and investigating the efficacy of the method on text collections.

Acknowledgements.

We thank Andrea Sahli for extensive help with the usability tests, Rashmi Sinha and Kevin Chen for helpful discussion, the San Francisco Fine Arts Museum for access to the dataset, and all our usability study participants. This research was funded by an NSF CAREER grant, NSF-IIS 9984741 and an IBM Faculty Fellowship.

REFERENCES

1. L. H. Armitage and P. G. B. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299, 1997.
2. M. L. Bernard. Examining the effects of hypertext shape on user performance. *Usability News*, 4(2), 2002.
3. P. Borland and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3):225–250, 1997.
4. C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*, 1999.
5. A. M. Elliott. *Computational Support for Sketching and Image Sorting During the Early Phase of Architectural Design*. Ph.d. dissertation, University of California, Berkeley, 2002.
6. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
7. S. R. Garber and M. B. Grunes. The art of search: A study of art directors. In *Proc. of CHI-92*, Monterey, CA, 1992.
8. M. Hearst, J. English, R. Sinha, K. Swearingen, and K.-P. Yee. Finding the flow in web site search. *Communications of the ACM*, 45(9), September 2002.
9. J. M. Jose, J. Furner, and D. J. Harper. Spatial querying for image retrieval: a user-oriented evaluation. In *Proceedings of ACM SIGIR '98*, pages 232–240, 1998.
10. M. Markkula and E. Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval*, 1:259–285, 2000.
11. J. Muramatsu and W. Pratt. Transparent queries: Investigating users' mental models of search engines. In *Research and Development in Information Retrieval*, pages 217–224, 2001.
12. W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, and C. Faloutsos. The qbic project: Querying images by content using color, texture, and shape. *SPIE: Storage and Retrieval for Image and Video Databases*, 1908, 1993.
13. M. Ortega, Y. Rui, K. Chakrabarti, S. Mehrotra, and T. S. Huang. Supporting similarity queries in MARS. In *ACM Multimedia*, pages 403–413, 1997.

14. C. Plaisant, B. Shneiderman, K. Doan, and T. Bruns. Interface and data architecture for query preview in networked information systems. *ACM Transactions on Information Systems*, 17(3):320–341, 1999.

15. K. Rodden, W. Basalaj, D. Sinclair, and K. R. Wood. Does organisation by similarity assist image browsing? In *Proceedings of ACM SIGCHI 2001*, pages 190–197, 2001.

16. B. Shneiderman, D. Byrd, and W. B. Croft. Sorting out searching: A user-interface framework for text searches. *Communications of the ACM*, 41(4):95–98, 1998.

17. R. K. Srihari, Z. Zhang, and A. Rao. Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval*, 2(2/3):245–275, 2000.

18. R. C. Veltkamp and M. Tanase. Content-Based Image Retrieval Systems: A Survey. Technical Report UU-CS-2000-34, Dept. of Computing Science, Utrecht University, 2000.