Photo Annotation on a Camera Phone

Anita Wilhelm[1], Yuri Takhteyev[1], Risto Sarvas[2], Nancy Van House[1], Marc Davis[1]

[1]School of Information Management and Systems
University of California Berkeley
102 South Hall, Berkeley, CA 94720-4600, USA
{awilhelm, yuri, vanhouse, marc}@ sims.berkeley.edu

[2]Helsinki Institute for Information Technology (HIIT)
PO Box 9800, 02015 HUT, Finland
risto.sarvas@hiit.fi

Abstract
In this paper we describe a system that allows users to annotate digital photos at the time of capture. The system uses camera phones with a lightweight client application and a server to store the images and metadata and assists the user in annotation on the camera phone by providing guesses about the content of the photos. By conducting user interface testing, surveys, and focus groups we were able to evaluate the usability of this system and motivations that will inform our development of future mobile media annotation applications. In this paper we present usability issues encountered in using a camera phone as an image annotation device immediately after image capture and users' responses to use of such a system.

Categories & Subject Descriptors: H.5.1 [Information interfaces and presentation (e.g., HCI)]: Multimedia; H.4.3 [Information systems applications]: Communications Applications; H.3.m [Information storage and retrieval]: Information Search and Retrieval

General Terms: Design, Human Factors

Keywords: Mobile Camera Phones, Automated Content Metadata, User Experience, User Motivation, Digital Image Management, Wireless Multimedia Applications

INTRODUCTION

With the number and adoption of consumer digital media capture devices increasing, more personal digital media is being produced, especially digital photos. As consumers produce more and more digital images, finding a specific image becomes more difficult. Often, images are effectively lost within thousands that are only demarcated by sequential file names. One solution to this image management problem is to enable users to create annotations of image content (i.e., "metadata" about media), therefore allowing consumers to find their photos by searching on information, instead of simply filenames.

Previous research in personal image management (surveyed in [5]) has facilitated annotation by using free-text, hierarchical and faceted metadata structures both textual [7] and iconic [1], drop down menus, drag and drop interfaces, and audio annotation with automated text transcription. Researchers have also sought to leverage the underlying temporal structure of photographed events to support browsing and retrieval. Consumer products are now beginning to appear which utilize metadata for image management, such as Adobe Photoshop Album 2.0, ACDSee, Apple iPhoto, and Adobe Photoshop CS. However, the vast majority of prior work on personal image management has assumed that image annotation occurs well after image capture *in a desktop context*. Time lag and context change then reduces the likelihood that users will perform the task, as well as their accurate recall of the content to be assigned to the photograph.

Mobile devices, however, are designed to take into account the users' physical environment and usage situations and can ultimately enable us to infer image content from mobile use context. Furthermore, by utilizing networked devices collaborative, co-operative applications are possible. If we can take advantage of the affordances of mobile imaging, we can overcome the loss of metadata in current digital photography due to time lapse and context change between image capture and image annotation, as well as use mobile contextual information to help to automate the image annotation process.

Networked mobile camera phones offer a good platform to apply these principles by providing us with a networked image capture device. While others have written about their effect on the content of photos (e.g., [3, 4]), we were interested in how they might be used to facilitate the annotation process. The purpose of our project was to create an infrastructure for networked cameras to allow users to assign metadata at the point of capture and to utilize a collaborative network, along with automatically captured environmental cues, to aide in automating the annotation process, thus reducing the effort required of the user.

METHODOLOGY

We built a framework ("MMM" for "Mobile Media Metadata") that enables image annotation at the point of capture using Nokia 3650 camera phones over the AT&T Wireless GSM/GPRS service [6]. We then gave 40 first year graduate students and 15 researchers camera phones to test our system for four months. We asked the students to brainstorm applications to build on top of this framework. Our evaluation consisted of three investigations. First, we performed user interface testing with five participants, giving them three scenarios each for phone use, videotaped their actions, and interviewed them afterwards about the use

CHI 2004, April 24–29, 2004, Vienna, Austria.
ACM 1-58113-703-6/04/0004.

scenarios and their current habits of image capture, storage, sharing, and retrieval. Second, all 55 participants were administered a weekly survey for seven weeks, inquiring about their use of the phones and the implemented image annotation system. Third, two focus groups discussed their image capture, storage, sharing, and retrieval habits. One group (eight subjects) consisted of users of this system and the other (seven subjects) was a general group of students. The former group additionally discussed their use of image annotation systems and this one in particular.

SYSTEM OVERVIEW

Utilizing the camera phone's hardware, network access, and software programmability, we built a client-server architecture. The client side software consisted of two components. The first component implemented the picture taking functionality and automatically gathered available contextual metadata before users uploaded the captured images to a remote server over the GPRS connection. The second component was the phone's built-in XHTML browser. It was used for all subsequent user interaction between the client and the remote server, which communicated with a collaborative repository of annotated images in order to help automate and facilitate the annotation process.

The first component, named Image-Gallery, was developed in co-operation with Futurice[1] for the Symbian 6.1 operating system on the phone. Image-Gallery automatically captured location metadata by storing the GSM network cell ID. Then, utilizing the username associated with each phone, it automatically captured the user's identification, as well as time and date at the moment of capture. This information was sent with the photograph, via the GSM/GPRS network to our server, where it was matched against a repository of annotated images. After Image-Gallery launched the phone's web browser, annotation "guesses" generated by a server-side program were returned to the user, through the XHTML browser on the phone to await confirmation or correction.

Keeping the human-in-the-loop, the XHTML browser presented the user with a series of screens suggesting metadata about the photograph. For each screen the server-side program would try to "guess" each answer by matching any previously submitted information against the collaborative repository of annotated images. The "guesses" were presented as drop down lists of prepopulated answers. The choice at the top of each list was deemed the most probable based on server-side matching algorithms. The user could then confirm the suggested annotation with one simple click, or correct the annotation by selecting a different option, including inputting new annotation text altogether (see Figure 1).

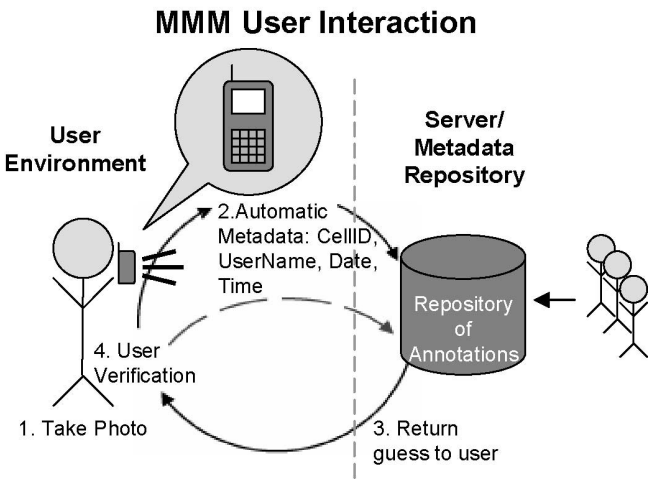

Figure 1. Mobile Media Metadata (MMM) User Interaction

The remote repository of metadata possessed a faceted hierarchical structure [1, 7]. By utilizing orthogonal hierarchies of descriptors that can be combined to make more complete descriptions, faceted classification structures enable rich description in ways that overcome the limitations of strictly hierarchical metadata structures and keyword based approaches used in most prior image annotation systems (e.g. FotoFile). One problem we faced, as we will discuss later, was the display and navigation of the faceted metadata structure on the limited screen size of the mobile phone.

USER INTERACTION CHALLENGES

By formally testing our infrastructure and by ongoing interaction with the 40 student subjects and 15 researchers we identified a set of challenges.

Network Unpredictability

The primary user difficulty was the unpredictability and limited availability of the GSM/GPRS network. The network often failed to transmit the image and/or metadata (both automated and user-assisted) to the server and was often very slow. Interactions that we had predicted would take 30 to 45 seconds per session often took 3 to 5 minutes. Users became very frustrated as the delay distracted their attention from their ongoing tasks and provided little feedback in the process. Users commented during testing:

"I have to keep staring at the screen to check for change even though I would rather pay attention to other things around me."

XHTML Browser Interaction

In effort to keep our prototype thin and simple [2], considering the large learning curve of the Symbian OS, we utilized the camera phone's XHTML browser as our principle user interface. The XHTML browser interaction, however, presented the users with interesting usability problems. Once the browser is launched, the form buttons

[1] http://www.futurice.fi

contained within the XHTML page do not correlate with the hardware buttons (called "softkeys") located on the phone client. Instead, these two softkeys, located just below the screen, contain hard-coded browser functions. To customize these softkeys, client-side programming is necessary. To avoid excessive client-side programming, all of our navigational options were contained within a form, inside the XHTML pages. Therefore, unlike full-client programs, like Futurice's Image-Gallery, which interact with the user by both softkeys as well as a central scroll key, all interaction within the browser (including all navigational and annotation options) necessary for our application were navigated by one central scroll key. Subsequently, the interaction followed one of a desktop web application more with the center key substituted for the mouse. This presented problems for our users:

"It was confusing to alternate from using the two big buttons under the screen (options) [i.e., softkeys] because once you are in MMM you should never use them or else you'll get kicked off site."

Metadata Hierarchy Display

Finally, the small screen of the mobile phone presented challenges for traversing a large facetted hierarchical metadata display. Neither the breadth nor depth of the classification structure could be displayed easily, as the former encountered screen real-estate limitations and the latter sequential page load latencies. We therefore decided to carefully select key nodes of the hierarchy and present the user with a limited user-focused hierarchy. This hierarchy contained approximately three depth levels across four facets (Person, Location, Object, and Activity). Each descriptor exposed in the interface was carefully selected to correlate with its node in the larger backend hierarchy. Therefore users could select one salient descriptor on the interface, but actually annotate many more. For example, the user could specify a location as "South Hall", but the image would be annotated as: *US> California> Alameda County> Berkeley> UC Berkeley> South Hall.* Any implied metadata stored in the backend hierarchy would automatically be added to the user's annotation.

This user-focused hierarchy was traversed as a series of screens containing drop-down lists. The users reported that they understood the interaction, however, because of the limited choices they were often unsure where to categorize their photograph.

Furthermore, because we allowed users to add new items to the hierarchy, the drop-down selections became very long. Users seemed to tolerate scrolling through 12-15 items, but once the list exceeded that length they complained that no "jumping" or "short-cut" mechanisms were available to help quickly traverse the long list. One user requested:

"I would like menus that wrapped or some ability to jump down on a menu (alphabetically?)"

Lessons Learned

Though we did in many ways reduce the cost of annotation to users by allowing them to remain in their environment while annotating, as well as reducing keystroke entry by allowing drop-down selection from a repository of inferred metadata, the network and user interface issues we encountered severely limited usability. Most participants only used the annotation system to complete required classroom assignments and even then, only annotated one to two facets per photo.

However, we feel that we did make some progress in reducing annotation cost, as users ultimately noted:

"For the most part, it's fairly easy to select items and click 'Next.' The annotation process isn't that hard."

Subsequently, we feel that by taking into account these learned future versions will only improve:

- Design for network unpredictability and errors. One possible solution is to limit continual network interaction by creating a full-client application. (The network should also improve over time.)
- Web applications run through the XHTML browser on mobile phones do not simulate full-client application navigation well. Use a prototyping methodology that simulates the specific mobile interaction better to test user experience.
- Presentation of a limited version of the faceted hierarchical structure must not be so limited as to overly constrain or confuse the user. Use of a different display metaphor, may be a better approach and lists of 12-15 items should not be exceeded.

USE PATTERNS AND MOTIVATIONS

To design systems to support image capture and re-use in general, to improve this infrastructure, and to create useful applications using it, we need to understand how camera phones affect users' photographic habits, as well as their motivations for annotation. We addressed these questions in our focus groups, interviews, and surveys.

Digital Images: The Funnel Effect

When we asked users about digital images in general, participants in the focus groups described a funnel effect in digital picture taking, sharing, and printing. They took many pictures, kept some of them, shared a selected group of those, and printed an even smaller subset. Digital imaging was for many a key element in the large volume of "throwaway" pictures, since, unlike "regular" photos, the marginal cost of each photo was zero. They reported that they would like to annotate only a subset of those taken, mostly only those good enough to share.

Additionally, student subjects were generally unconcerned about metadata for future use: for example, identifying people in photos—they said that they already knew these people. We suspect that this short-term perspective may be

due in part to the relative youth and childless status of most of our subjects. They did not seem concerned, for example, with sharing images with future generations.

Camera Phone Photos: The Power of Now

Because users were able to carry their camera phones much of the time, they reported taking more humorous or ad-hoc images than they would with their "normal" cameras, which they often only carried to specific events or for specific purposes. One user reported taking an image of a rather sad (droopy) palm tree outside of the school building because he wanted to capture its melancholy that day; another took pictures of students filling a water fountain with bubble bath. In such cases, the "power of now" was apparent. They were able to capture unique or funny moments in their daily lives and communicate them to others via images. This is consistent with other researchers' findings that people take different kinds of photos with camera phones [3, 4].

Like Ito and Okabe's [3] users, our users reported a short-term orientation toward the photos taken with the camera phones, with more interest in sharing than in searching or retrieving their photos. One group wrote a script that would automatically publish selected photos to a personal web page, with an attached caption (moblogging). Other users shared their photos by using the imaging device itself: showing people images on the camera phone. Still others used email, Bluetooth, and infrared capabilities to share images with others. Near the end of the semester, we supplied the students with a web-based browsing tool to view images and their annotations via the desktop. Users reported that this made sharing easier and added great value to the application. Furthermore, users preferred searching and browsing based on input metadata via the desktop rather than the phones and suggested further desktop-based annotation capabilities.

Selective Metadata Annotation

To our surprise, our subjects were generally not interested in fully annotating photos by keywords. They simply wanted to attach one or two salient identifiers. Annotations often took the form of captions rather than standard metadata: the reason why they took the picture, a witty remark, or something personal shared with the observer. This was true of photos taken with the camera phones and other cameras, but the immediacy of camera phone photos seemed particularly well-suited to this kind of annotation.

Lessons Learned

From the above, we conclude:

- Mobile camera phone use highlights the "power of now" in always being available for ad-hoc picture taking.
- For our camera phone users, sharing and browsing are more important than searching or retrieval.
- A desktop component adds great value to the mobile application by easing search, sharing, and quick browsing.
- User preferences for annotation are generally limited to a few favored images, and some key information for each photograph.

CONCLUSIONS

From this group of users, we conclude that mobile camera phones enable a new approach to annotating media that can reduce user effort by (1) facilitating metadata capture at the time of image capture, (2) adding some metadata automatically, and (3) leveraging networked collaborative metadata resources. As networks improve, our problems with network latency and unreliability will be reduced. However, user interface and system designs for mobile image annotation need to overcome the challenges of text entry and hierarchical display and navigation on mobile devices. We also need to develop hybrid solutions that integrate desktop and mobile application components into more complete and appropriate solutions than either can offer alone.

More generally, we need to understand and design for the emergent behavior resulting from changes in technology. Digital imaging, in general, and camera phones in particular, make new kinds of imaging behavior possible. The ready availability (and current low image quality) of camera phones encourages the capture of images for short-term uses affecting the kind of annotation currently desired. As image quality improves, we expect that users will add to these ad hoc uses more traditional (long-term) imaging behavior with more need for metadata.

REFERENCES

1. Davis, M. Media Streams: An Iconic Visual Language for Video Representation. in Baecker, R.M., Grudin, J., Buxton, W.A.S. and Greenberg, S. eds. Readings in Human-Computer Interaction: Toward the Year 2000, Morgan Kaufmann, San Francisco (1995) 854-866.
2. Edwards, K.W., Bellotti, V., Dey, A.K., Newman, and M.W., Stuck in the Middle: The Challenges of User-Centered Design and Evaluation for Infrastructure. Proc. CHI2003. AMC Press, 297-304.
3. Ito, M. and Okabe, D., "Camera phones changing the definition of picture-worthy," Japan Media Review (2003). http://www.ojr.org/japan/wireless/1062208524.php
4. Koskinen, I., Kurvinen, E. and Lehtonen, T.-K. Professional Mobile Image. IT Press, Helsinki (2002).
5. Rodden, K. and Wood, K.R., How Do People Manage Their Digital Photographs? Proc. CHI2003, ACM Press (2003), 409-416.
6. Sarvas, R., Herrarte, E., Wilhelm, A., and Davis, M., Metadata Creation System for Mobile Images. Proc. MobiSys2004, ACM Press (Forthcoming, 2004).
7. Yee, P., Swearingen, K., Li, K. and Hearst, M., Faceted Metadata for Image Search and Browsing. Proc. CHI2003, ACM Press (2003), 401-408.