Metadata Creation System for Mobile Images

Risto Sarvas
Helsinki Institute for Information Technology (HIIT)
P.O.Box 9800,
02015 HUT, Finland
+358 9 694 9768
risto.sarvas@hiit.fi

Erick Herrarte, Anita Wilhelm, Marc Davis
Garage Cinema Research Group, School of Information Management and Systems, UC Berkeley
102 South Hall, Berkeley, CA 94720, USA
+1 510 642 1464
{herrarte, awilhelm, marc}@sims.berkeley.edu

ABSTRACT
The amount of personal digital media is increasing, and managing it has become a pressing problem. Effective management of media content is not possible without content-related metadata. In this paper we describe a content metadata creation process for images taken with a mobile phone. The design goals were to automate the creation of image content metadata by leveraging automatically available contextual metadata on the mobile phone, to use similarity processing algorithms for reusing shared metadata and images on a remote server, and to interact with the mobile phone user during image capture to confirm and augment the system supplied metadata. We built a prototype system to evaluate the designed metadata creation process. The main findings were that the creation process could be implemented with current technology and it facilitated the creation of semantic metadata at the time of image capture.

Categories and 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, Content-based Image Retrieval, Digital Image Management, Wireless Multimedia Applications

1. INTRODUCTION
The amount of personal digital media is increasing. Consumer devices such as camera phones greatly facilitate the daily creation of digital images, audio, and video. However, this brings with it the inherent problem of media management: managing large amounts of personal digital media based on their content is time consuming and difficult. Human memory and browsing can manage the content of ten or even a hundred pictures; when one has thousands of pictures, the ten pictures one is looking for are effectively lost. Management of personal digital media must be automated. However, effective organizing, searching, and browsing of digital media based on their content is currently not possible in most consumer media management tools.

One solution for automating media management is to have information about the media's content in a computer-readable form – media content metadata. Content metadata describe what is in the media, for example, who are the people in the picture, where was a scene in a movie shot, or what concert is this audio track from. Also, it can describe the semantic structures in media, like who is doing what to whom in a picture. The latest consumer photo management software programs and research systems are increasingly based on creating and using metadata about image content [2, 3, 11, 12, 18].

The past decade of multimedia content processing research has focused on automatically extracting metadata from digital media (see, *e.g.*, [1, 5]). However, this kind of metadata is relatively low-level, not necessarily the kind of semantic metadata people are looking for. On the other hand, annotating high-level semantic metadata manually is too time-consuming for most applications and users. This is referred to as the semantic gap between the low-level metadata acquired automatically and the high-level metadata people are looking for [9, 16]. To bridge this gap the automatic extraction algorithms must be combined with the human ability to easily identify semantic information [7, 8].

Related work of [18] takes advantage of GPS-derived location metadata as the basis for organizing and sharing images. In [13], researchers built a system that leverages location metadata to infer image content. In [20] the authors describe a semi-automatic image annotation process that combines content-based image retrieval and user verification to achieve correct high-level metadata. In [19] mobile phone cameras and automatically available contextual metadata are combined with an annotation interface to generate metadata. Our research, however, goes beyond all these approaches in a novel automated metadata creation process which leverages three types of contextual metadata (spatial, temporal, and social, *i.e.*, "where", "when", and "who"), the efficiency of automatic metadata inferring, sharing of media and metadata, and human-in-the-loop metadata verification.

Semantic information should be gathered at the time of capture when the context of capture and human agents who can help disambiguate and describe that context are still available,

MobiSYS'04, June 6-9, 2004, Boston, Massachusetts, USA.

otherwise semantic metadata is often not created or it is lost either because of lack of integration in the media production process or simply because people can no longer recall information that would have been available at the time of capture. This is true of both professional production and consumer media production. In professional production like motion pictures, very little information about the media's content survives the production process (*e.g.*, who is in a scene, when does a scene begin and end, or where is it taking place). The same is true of consumer media production, where people, places, and events obvious at the time of capture, usually fade from people's memory. Therefore, the ideal approach is to acquire the metadata information when it is available: both the information in the minds of people at the time of capture, and the information in the technical devices the people are using at the time of capture (*e.g.*, camera, camcorder, calendar, phone, PDA, game deck, etc.).

However, acquiring semantic information at the time of capture is problematic. The user might not be willing or have time to annotate an image with metadata (*i.e.*, inputting information about an image's content), and the technical devices and tools available might have no means of sharing their information at the time of capture.

As a media recording device, current mobile phones with cameras have key features that make them a promising metadata annotation and sharing platform. Unlike consumer media recording devices like digital cameras or camcorders, mobile phones have computing power that can be accessed via open and standardized programming interfaces (*e.g.*, J2ME, BREW). Mobile phones also have inherent network connectivity, the user interfaces are developing at a rapid pace, and the location awareness of mobile phones is getting more accurate.

1.1 Research Problem and Solution

Automatic media management requires semantic metadata about media content. However, computational extraction of semantic content metadata from media is not easy. Much of semantic content information is available at the time of capture, mostly from people, but also from devices and tools used in production. This information is effectively lost as time goes by.

Rather than attempting to reconstruct semantic metadata by analyzing media long after it has been captured, we seek to leverage the spatio-temporal context and social community of capture to computationally assist users in creating useful semantic metadata about media content at the time of capture.

In this paper we describe how we create, share, and reuse semantic metadata at the point of capture using a mobile phone camera. We address the problem of creating semantic content metadata by designing the metadata creation process based on the following four principles:

1) Gather all automatically available contextual metadata at the time of capture.
2) Use metadata and media similarity processing algorithms to infer and generate new metadata for captured media.
3) Share and reuse media and metadata among users to facilitate metadata creation and new media applications.
4) Interact with the user during capture to confirm and augment system supplied metadata.

We also describe a prototype system we implemented to test the automated metadata creation process called "MMM" ("Mobile Media Metadata"). In the MMM prototype we used Nokia 3650 GSM camera phones to take images and for user interaction, and a HTTP protocol over a GPRS network to communicate with a remote application server for metadata processing and storage (see Figure 1).

Section 2 describes in detail the designed metadata creation process. Section 3 presents the implemented prototype system and its main components. Section 4 presents experimental results from using the system, and system evaluation. Section 5 discusses future research, and Section 6 concludes.

2. METADATA CREATION PROCESS

In this section we first discuss the motivation for annotating media in general, and give an overview of the designed process. After that we describe the four above-mentioned principles in more detail. For each of the four principles we discuss the design rationale, theory, and how we implemented it in our prototype. A more technical description of the prototype system is in Section 3.

2.1 Motivation and Overview

The motivation for creating media content metadata is the facilitation of media management and the enabling of potential media applications that make use of that metadata.

A repository of annotated media would make searching, browsing, and managing media significantly easier. Content metadata would enable structured and semantically rich queries. For example, it would be possible to search or browse images that are taken in a certain location, at a given time, and with certain people in the image. Furthermore, the query could take into account who took the picture, what is happening in the picture, and so on. The annotated media repository could be shared, leveraging the information created by anyone.

The mobile media metadata technology we describe in this paper could enable a wide variety of mobile media products and services for mobile creation, sharing, and (re)use of media and metadata. Some examples include: networked photo albums, personalized media messages, matchmaking services, mobile wish

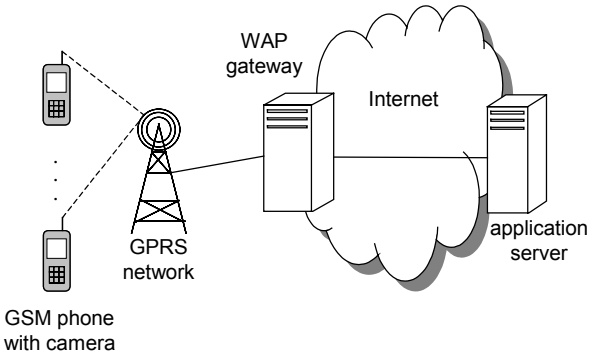

Figure 1. The network architecture of the MMM protoype where the mobile phones were connected to a remote server on the Internet.

lists and shopping, and travel information.

2.1.1 Overview of the Process

Our designed metadata creation process has five main parts (see Figure 2). The first part of the process is the image capture itself done by the user using the camera phone. This starts the actual metadata creation process. The following parts of the process leverage the four principles mentioned above.

Right after the image has been captured the next step gathers contextual metadata available to the system. The following third step consists of two parts: metadata and media similarity processing, and metadata and media sharing and reuse. After the similarity processing, sharing, and reuse, the system generated metadata is presented to the user for verification. The user verified metadata is then sent back to the previous part for more processing. This loop continues until the user ends it, or the system generates no new metadata.

2.1.2 Use Scenario: The Campanile

This use scenario describes how the designed metadata creation process should work in practice. The scenario is from the user's point of view and refers to Figure 2. The described scenario is an ideal use case to describe the main concepts of the process. Sections below describe what parts of it were implemented in the prototype system.

The user is visiting the UC Berkeley campus, and takes a picture of the Campanile tower, which is a very popular object of photography in Berkeley. She points the camera phone at the Campanile, and pushes the capture button. As the image is captured by the camera phone, the device also stores the contextual metadata available to it: location data, time, date, and user name.

The image and the contextual metadata are uploaded to a remote application server. The server processes the image and the contextual metadata to find similarities with other images and metadata. The similarity processing takes advantage of all the users' images and metadata available to it. In other words, it shares all the images and metadata among the users of the system, and reuses it. Because the Campanile is both a popular object of photography and an easily distinguishable object, there is a large number of similar images taken at that location.

Therefore, based on the location data (*e.g.*, EOTD location information, or a GSM network cellID), and image pixel data (blue upper-part is the sky, green lower-part is the trees, and a light gray vertical rectangle in the middle is the tower) the server can find similar images and metadata. Based on those images and metadata, in addition to simple processing, the server comes up with the following metadata and an accuracy estimate for each metadata: location city is Berkeley (100%), day of the week is Thursday (100%), object in the picture is Campanile (62%), and the setting of the picture is outside (82%).

The metadata which the system is not hundred percent sure about is sent back to the user for verification. The user verifies that the picture is of the Campanile and that the setting is outside by a push of a button. This is sent back to the server, and once it is received by the server, the server can both add new information automatically to the image (*e.g.*, location metadata like city, county, state, country), and also suggest additional metadata that the user can approve or not (*i.e.*, continue the metadata generation and user verification loop).

For example, after the user has verified that the image is of the Campanile the system can add more information based on that verification: the location of the picture is the UC Berkeley campus, City of Berkeley, County of Alameda, State of California, U.S.A.; the image was taken on Thursday, October 9th, 2003, 15:34 PST; the object of the picture is The Campanile also known as the Sather Tower, a stone bell tower, completed in 1914, height 94 meters; and so on. By one push of a button the user has annotated the image with almost twenty pieces of metadata.

2.2 Gathering of Contextual Metadata

The first step after image capture, and the first of our four design principles is the gathering of automatically available contextual metadata at the time of capture. The metadata can be derived from

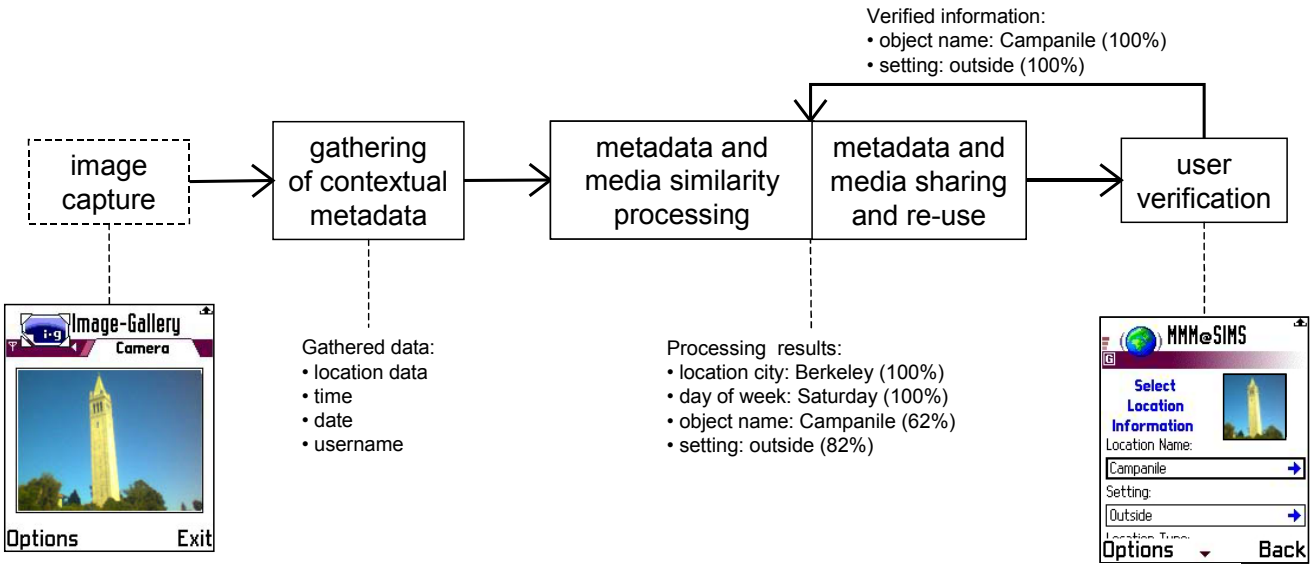

Figure 2. The MMM metadata creation process.

the media recording device itself or from the network if the device is connected to one. The contextual metadata can be information about the computing context (*e.g.*, network resources, nearby devices), spatial context (*e.g.*, cellID or GPS coordinates), temporal context (*e.g.*, time, date), social context (*e.g.*, user name, user profile, presence of other users), and environmental context (*e.g.*, temperature, lighting) [6, 15].

Most of the easily available contextual metadata is semantically low-level information that does not say much to a person about image contents. For example, location coordinates can be said to be very detailed information but seldom mean anything as such to a person. People use location information at a higher semantic level, like the name of a building, street address, or a geographic location. Nevertheless, the contextual metadata is valuable as basic unambiguous metadata to be processed further to derive and infer higher level media and context semantics.

However, contextual metadata is not often automatically available. Different media recording devices have different capabilities to provide or access context information. Most media recording devices can provide basic temporal information like time, date, and duration of recording. Other information that is often available is, for example, camera's properties at the time of capture (*i.e.*, camera model, exposure time, aperture, and so on). Spatial information is rarely available. For example, few consumer digital cameras or camcorders have access to any location information at the time of capture. Also, few media recording devices have any information about the user of the device or the user's behavior or preferences.

In other words, the main problem in gathering contextual metadata is that usually the metadata is not available at all – the device does not have the information.

The other problem is sharing the information among devices (*e.g.*, other recording devices, servers, or media repositories) so that it can be used in metadata creation. Especially at the time of capture, other devices in a network might have information about the current context, more processing power, or better access to relevant information. For example, most consumer digital cameras and GPS devices have no standardized way of communicating between each other. So the other big problem in gathering contextual metadata is the lack of network connectivity and pre-defined protocols and standards between devices.

2.2.1 Implementation

In the scenario and the MMM prototype we used a camera phone as the media recording device. Unlike most digital cameras and camcorders, current camera phones, in addition to media recording capabilities, have an inherent network connection (*e.g.*, GPRS) and a wide range of basic contextual information available: location (*e.g.*, network cellID data, or EOTD data), time, date, and user identification (*e.g.*, SIM card identification). Also, current mobile phones have programming APIs (*e.g.*, J2ME and Symbian OS) that enable configuration and programming of the phone to better suit contextual metadata gathering. To the best of our knowledge, practically none of these functionalities or data is available in most other consumer media recording devices which makes camera phones an important emerging platform for digital media and metadata creation, sharing, and reuse.

The programming APIs on mobile phones enable access to more than just the basic contextual information of time and space. Mobile phones usually have calendars, address books, and usage logs in them. Some mobile phones also have Bluetooth connectivity and can get information about all other Bluetooth devices within range. This kind of information can be valuable metadata for further processing. For example, if the calendar shows a meeting with a colleague at the time of capture, and the colleague's mobile phone is within the Bluetooth network's range, it is highly likely that the colleague is somehow associated with the captured image.

Our MMM prototype implementation has a C++ program on the phone's Symbian operating system that the user uses for taking the images and uploading them to a server. The program automatically gathers the network cellID, the capture time and date, and the user name. When the user chooses to upload an image, our client-side application sends it and the contextual metadata to our remote server over the GPRS network. A more detailed description of the prototype is provided in Section 3.

2.3 Metadata and Media Similarity Processing

The goal of media and metadata similarity processing is to find similar media and metadata, and reuse their metadata. Metadata similarity compares the metadata of a captured image to already existing images, and if they have, for example, the same location city, then the broader geographic metadata can be automatically reused (*e.g.*, county, state, country). Metadata processing can also take advantage of regularity in media capture contexts. For example, regularity based on the simple assumptions that people tend to photograph the same people (*e.g.*, friends, family), and that locations where people usually take pictures also have regularity (*e.g.*, home, work, or hobby).

Secondly, the similarity processing can have simple algorithms for inferences like what day of the week a given date is, or lookup functions such as to what city does the given location data map.

Thirdly, media similarity processing can take advantage of content-based image retrieval algorithms that take the image pixel data as input and produce information about an image as an output, for example, color histograms, textures, face detection, or shape matching (see, *e.g.*, [4, 17]) which can be used to measure media similarity.

Each of these approaches can also have an accuracy estimate that gives a probability whether the inferred metadata is correct. For example, an image processing algorithm can be 64% sure that the image is taken outdoors, a metadata similarity algorithm can be 84% sure that because the location and person metadata are the same, the event metadata is also, and a simple algorithm can be 100% sure that October 9th, 2003 is a Thursday.

Combining these three approaches of similarity processing (metadata processing, simple inferences, and media processing) provides the most accurate results. For example, after receiving the picture of the Campanile from the phone, image similarity processing could come up with similar images and most of them are of the Washington Monument, and some of them are of the Campanile in Berkeley. Taking advantage of the location data sent with the image, a simple lookup algorithm could quite accurately determine that the image was taken of the Campanile in Berkeley, rather than the Washington Monument. Now the metadata similarity algorithm can compare the metadata of all the

images of the Campanile, and reuse the information about its detailed location, height, building type, and so on.

2.3.1 Implementation

In the MMM prototype we implemented metadata processing as a separate library of algorithms that could be added, edited, and deleted. We implemented two metadata inferring algorithms: a location guessing algorithm and a person-guessing algorithm. In our next version of the MMM system we plan to integrate media similarity processing algorithms [17] and interleave media and metadata similarity processing.

The implemented location-guessing algorithm compares the captured image with other images. The algorithm compares the new photo to other pictures taken by the same user, or other pictures where the user is in the picture to infer semantic location from cellID information. It also looks for similarities in time, date, and day of the week. For example, if the difference in time between two pictures taken by the same user is relatively small then the location metadata is more probably the same. The algorithm combines all this information, and the end result is a list of locations, sorted by their accuracy estimate.

The person-guessing algorithm guesses the person in the picture. The algorithm finds how many times the user who took the picture is related to other images, and their owners or subjects. For example, if the user who took a picture is often the subject of someone else's pictures, the algorithm increases the probability that that someone else is the subject of this newly captured image. The algorithm takes also into account similarities in time and space. The algorithm runs through these functions, and a person who gets most points from the tests is the most probable person in the newly captured image.

2.4 Metadata and Media Sharing and Reuse

One of the main design principles in our metadata creation process is to have the metadata shared and reused among all users of the system. This means that when processing the media and metadata, the system has access not only to the media and metadata the user has created before, but the media and metadata everyone else using the system has created. Although this is part of the media and metadata similarity processing, we bring it forth as a separate and significant part of the overall process.

As mentioned in the previous subsection, one of the main benefits of sharing media and metadata among users is the possibility to reuse the metadata information (*i.e.*, copy the existing metadata). By reusing metadata, the metadata creation process can be significantly automated, meaning less user's time and effort spent in annotating. Only one person needs to add the metadata and then it is automatically available for the benefit of everyone else.

The other benefit of sharing metadata is having a common metadata format for different people's image annotations. Emerging standards like MPEG-7 are seeking to create a common infrastructure for media metadata creation and sharing. As the metadata format is shared and common among the users, people can exchange or share images with the accompanying metadata information, and the metadata is usable by others.

For example, popular locations where people take pictures of the same people, objects, or events can have a great amount of metadata readily available, like in the Campanile use scenario or in a user's family home.

In addition to the more technical benefits of facilitating the metadata creation process, sharing metadata is an opportunity for building social and communal relations between people. People with similar interests, or connected acquaintances could find each other. For example, an ornithologist who photographs birds could find out who is the person who is also annotating the local bird species.

One of the issues in sharing information between users is privacy. People take personal pictures and provide the system with personal information. If people share the information, an important set of questions arise as to how to protect user privacy without hindering the benefits of sharing metadata. While a detailed discussion of this issue is beyond the scope of this paper we did find that in addition to standard ways of protecting privacy through anonymization, an important approach is to understand the tradeoffs users are willing to make in exchanging some amount of privacy for a perceivable benefit. Possible perceive benefits of metadata sharing (*e.g.*, easier management and search of media) may offset some concerns about media and metadata privacy. This is a topic for further study.

Another problem with sharing metadata is the problem of language and vocabulary: different people call the same things by different names, and use the same name for different things [10]. Inconsistent vocabulary among users reduces the benefits of sharing. If people use incommensurate descriptions for the same things, the metadata information is not reusable. This problem was one of the research questions we explored in our implemented prototype. We investigated such questions as how would users utilize a shared semantic description scheme for annotating personal media? Also, if we allow free creation of semantic descriptors by diverse users, how divergent would the descriptor set become?

2.4.1 Implementation

In the implemented prototype we shared all the images and metadata among all users. The issues of privacy and ownership were left outside the scope of the prototype.

The vocabulary problem was addressed by having a predefined metadata structure. The structure itself was predefined (*i.e.*, the relations, hierarchy, categories, and so on), but to enable the users to input more metadata to the system for other people to use, the values of the structure could be edited. For example, a user could add the first name and the last name of the person in an image, but could not add a new category "nickname" into the structure. The metadata structure is described in more detail in Section 3.

2.5 User verification

As mentioned above, the metadata processing algorithms give estimates on the accuracy of the metadata they generate. In other words, they make educated guesses with probability estimates.

The purpose of the fourth step in our process is to get the user to verify the information. If the algorithms have guessed correctly, the user can easily confirm the guesses, and the verified information can then be used as input for the next cycle in the loop. The verification can be made easier by providing the user with choices where the most probable choice is presented first.

For example, in the Campanile use scenario, the processing could have been totally wrong: Instead of taking a picture of the Campanile bell tower outdoors, the user might have taken a

picture of a gray bottle, the background happened to be blue, and the location was indoors, two hundred meters south of the Campanile. The image processing would have found similar images, and probably most of them would have been of the Campanile, because the location data was so close. Therefore, the user must be asked to verify the guesses.

From the system's point of view the user is the best source for getting semantic information about the captured image. Unlike the media and metadata processing algorithms, people can easily identify other people, locations, events, and objects in pictures. They can easily describe semantically complex relations. Of course, people make mistakes (*e.g.*, remember people's names incorrectly, make spelling errors, or even intentionally annotate misleadingly), therefore, human error should be considered in system design.

For the purpose of verifying the metadata, the user is an exceptionally valuable part of the process. However, two issues rise from this user interaction: the user's motivation to spend time annotating and the usability of the interaction.

There is the possibility that the user is not motivated to annotate the picture, in other words, to verify the system supplied metadata. Often consumer photography involves taking quick snapshots rather than spending a lot of time positioning the camera, adjusting lighting, or choosing filters. This implies that consumers do not spend lot of time per image taken. Therefore, asking a user to annotate each and every picture taken immediately after capture can be too time-consuming in some situations.

The user's motivation for annotation can be affected by minimizing the effort required by either automating it as much as possible, or by making the annotation process entertaining. Also, depending on the application of the metadata, the user might be motivated if the effort of annotating now is offset by clear subsequent benefits. The usability of the interaction is related to the motivation. If the verification is difficult or slow to use, the user is less willing to provide the information.

Once the user has verified the metadata to be either right or wrong, the information is sent back to metadata processing. If the metadata was right, it can be used as an input for further media and metadata similarity processing. If it was wrong, and there were no correct options for the user to choose, the media and metadata processing takes that into account, and presents other options or a possibility for the user to provide that information.

2.5.1 Implementation

The user verification process, or annotation process is implemented in the MMM prototype by using the camera phone's XHMTL browser to communicate between the user and the remote server. On the server side, there is a program that handles the user interaction, running of the similarity processing algorithms, and XHTML page generation. The interaction is a form-based dialog where the server program creates the page using the metadata it wants to be confirmed, and the user either confirms the supplied metadata or selects other metadata and sends their choices back to the server.

To automate the user's annotation effort, the server tries to guess the metadata as much as possible, and leverage previously annotated images. This is the purpose of the similarity processing and metadata sharing and reusing parts of the process described above.

3. PROTOTYPE DESCRIPTION

To test the designed metadata creation process and to identify potential problems, we implemented a prototype infrastructure called "MMM" for "Mobile Media Metadata" based on the process described above.

The MMM prototype implementation was a client-server model where the clients are the camera phones connected to the Internet via a GPRS network using the WAP protocol, and the server is a web server connected to the Internet. The objective of the prototype was to understand better the technological challenges in implementing the metadata creation process on mobile phones and to provide a testing environment for user studies of media and metadata creation, sharing, and reuse.

The architecture, depicted in Figure 3, consists of seven main components and a metadata structure used for describing image content. The main components of the MMM prototype are as follows:

- ***Image-Gallery***: A client-side application for taking images, gathering contextual metadata, uploading the image and metadata, and initiating the user annotation.
- **XHTML Browser**: A client-side browser for the annotation interaction implemented in XHMTL forms connected to the web server.
- **Upload Gateway**: A server-side gateway that handles the communication between *Image-Gallery* and the server, and

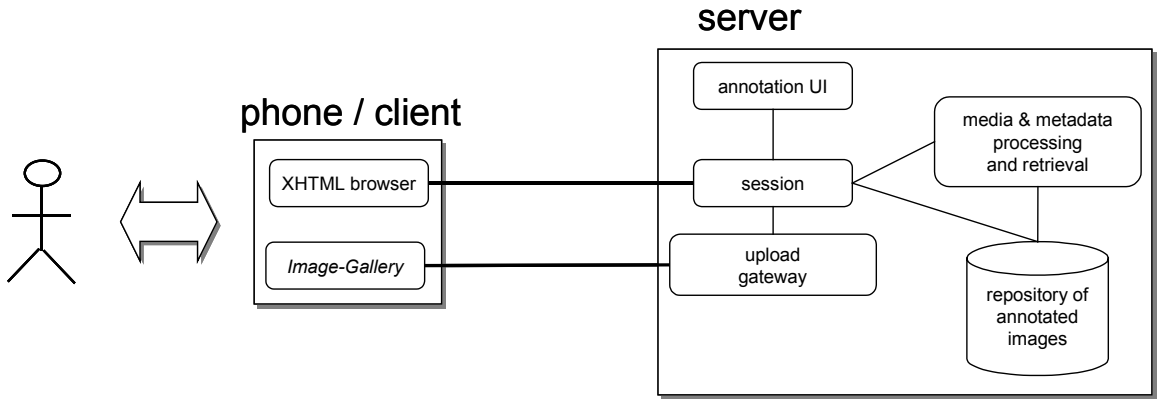

Figure 3. The implemented MMM prototype system architecture and its main components.

receives the uploaded image and metadata. The Upload Gateway implements the Nokia Image Upload Server API version 1.1 [14].

- **Session Class**: A server-side session object, which is instantiated for every user image annotation session. It manages the annotation process, and communication to the metadata processing component and the repository of annotated media. It maintains the state of the annotation process for the uploaded image and provides the application logic/behavior.
- **Media & Metadata Processing and Retrieval Module**: A module for automatically identifying metadata and associated probability estimates for the created metadata. This module consists of two algorithms that retrieve metadata based on previous user annotated images, metadata annotation frequencies, and spatial, temporal, and social information.
- **Annotation UI Module**: A user interface module that encapsulates and generates XHTML forms for the user. It also parses and verifies returned values.
- **Repository of Annotated Images**: A repository where both the metadata and images are stored.

The following sections first describe the metadata structure used, then each of the main system components and the annotation process in detail.

3.1 Metadata Structure

In describing the contents of an image we used a simplified faceted metadata hierarchy. The structure was based on the faceted metadata hierarchy designed for Media Streams, an iconic visual language for media annotation, retrieval, and repurposing [8]. The structure has main categories called facets. The objective of these facets is to be as independent of each other as possible, in other words, one can be described without affecting the others. In our structure the facets were Person, Location, Object, and Activity. Faceted metadata has many advantages for annotating, searching, and browsing media [22].

Each facet contained a tree hierarchy of more detailed subcategories. For example, the Object facet contained subcategories of types of objects: Animals, Architectural Objects, Clothing, Containers, Food, Furniture, Natural Objects, Plants, Play Things, Symbolic Objects Tools and Vehicles. The nodes of the tree structure could contain more nodes, or leaves which would be the final values either selected or typed in by the user.

Facets can also contain subfacets to greatly improve the expressive power of a small descriptor set through faceted combination. For example, the Location facet has a Geographic subfacet, a Type subfacet, and a Setting subfacet which can be combined to make many different, but semantically related, location descriptions:

- Geographic Location: USA > California > Berkeley > UC Berkeley > South Hall
- Location Type: Building
- Location Setting: *Inside*
- Geographic Location: USA > California > Berkeley > UC Berkeley > South Hall
- Location Type: Building
- Location Setting: *Outside*

An image in our metadata structure can have one facet as primary, called the Main Subject. This indicates what an image is primarily about. In addition to the Main Subject, an image can have several additional facet annotations, which are not primary. For example, a picture of a bus in Berkeley (see Figure 4) could have at least four top-level facets annotations: two Objects (one of them selected as the main subject by the user), Location, and Activity. The simple bucketizing of photo content into people, locations, objects, and activities can greatly aid the metadata and media processing algorithms in determining which prior media and metadata to compare a new image to.

3.2 Implemented Annotation Process

The metadata creation process, depicted in a sequence diagram in Appendix 1, starts by the user capturing an image using the *Image-Gallery* program on the phone. Once the image has been captured, the program asks the user to select the main subject of the image. The program also saves the contextual metadata available at the time of capture (*i.e.*, time, date, user name, and network cellID).

After taking the image the user uploads the image to a server using the same *Image-Gallery* program. On the server side the upload gateway component receives the login from the *Image-Gallery* program, and creates a new session for handling the user interaction. The new session is uniquely identified by the system using a session id that is generated by the system. From the created session the upload gateway sends the session id to the *Image-Gallery* application on the phone. *Image-Gallery* uploads

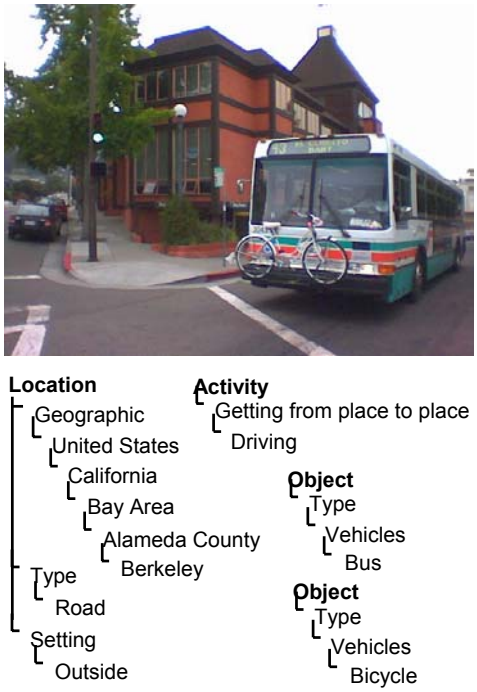

Figure 4. An example image annotation.

both the gathered contextual metadata and the image to the upload gateway. The upload gateway passes on the image and the metadata to the session object. The session object stores and associates the uploaded image and metadata in the repository.

After successful upload, *Image-Gallery* launches the phone's XHTML browser to a generated URL, which consists of the web server URL and the session id. The web server parses the HTTP request, the session id is verified, and the session object associated with that session id is retrieved.

The session calls the metadata processing and retrieval module to run its algorithms on the image and the metadata of the session. The metadata processing and retrieval module accesses the repository of annotated images according to its algorithms, and finally returns the inferences it has created. It returns new metadata and a probability estimate for each metadata. From this information the session generates guesses for the user, and these guesses are sent to the annotation UI module, which creates an XHTML form from the guesses, and inserts the required user interface components. This form is then returned to the session, and the session returns it back to the phone's browser. From the phone's browser the user can verify the information presented by pushing the submit button on the form. This information is then passed back to the session.

The session calls the metadata processing and retrieval algorithms based on this verified information. The metadata processing and retrieval algorithms once again access the repository of annotated media and return new metadata inferences along with a probability estimate for each inference. From this information the session composes new guesses to be presented to the user for verification. Then it calls the annotation UI module to generate an XHTML form from the guesses. This form is then passed to the phone's browser. From the phone's browser the user can once again verify the guesses.

If the user does not verify the information but exits the annotation process, this information is passed to the session, which stores the created metadata to the repository and terminates itself.

The annotation interaction between the system and the user is an iterative process consisting of the system generating guesses in the XHTML forms and the user verifying the guesses (*i.e.*, generating correct metadata). This annotation interaction loop continues as long as the system generates new guesses or until the user terminates the interaction.

3.3 Client-Side Components

The main components on the client-side are the client itself, that is the phone, the *Image-Gallery* program, and the XHTML browser. See Figure 1 for a network architecture, Figure 3 for a conceptual system architecture, and Appendix 1 for a sequence diagram and screenshots.

3.3.1 Mobile Phone with a Camera

The phones in our prototype are Nokia 3650 GSM phones[1] that have a built-in camera for recording video and taking still images. The model also has a GPRS data network connection. The HTTP protocol supported by the applications and the network were used to connect to the remote server on the Internet. The protocols were provided by the Symbian 6.1 operating system[2] installed on the phone. The application development on the phone was done using the Symbian C++ API provided in the Symbian OS SDK in order to take advantage of these protocols. The GSM/GPRS network used was the AT&T Wireless[3] network in the San Francisco Bay Area.

3.3.2 Image-Gallery

The program for taking images and uploading them was a C++ program for the Symbian OS named *Image-Gallery* (co-developed with Futurice[4], and configured for this particular prototype). The uploading of the image and metadata was done using the HTTP protocol implementing the Nokia Image Upload Server API version 1.1 [14].

3.3.3 XHTML Browser

The XHTML browser used for the annotation process was the default browser on the Nokia 3650 phone. The browser supports XHTML Mobile Profile [21].

3.4 Server-Side Components

The server-side components are all located on the same physical computer. The computer has an Apache HTTP server[5] and Apache Tomcat as the Java servlet container[6]. The following subsections describe the main Java modules (*i.e.*, the main server-side components) in the MMM prototype.

3.4.1 Upload Gateway

The upload gateway component consists of a set of Java servlets and Java objects that implement the Nokia Image Upload Server API version 1.1 [14]. The upload gateway acts as the interface between the *Image-Gallery* application running on the phone and the server side components of the system. It abstracts away the image and metadata transfer between these components from the rest of the system by encapsulating the transfer protocol.

The functionality implemented in the upload gateway is connection authentication using a user id and a password, *Image-Gallery* application interaction and data transfer between the client and the server (excluding the XHTML communication), and the initiation of a session object for image annotation. The upload gateway is also responsible for communicating the session identification number to the client to use during the user interaction process.

3.4.2 Session

The Java session class manages and handles the user interaction process. Each user interaction session has its own instance of the session class identified by a unique session identification number. The session object manages both the state of the user's annotation process and the user's technical preferences (*i.e.*, the phone model and web browser used). The user's annotation state includes the latest user interface (*i.e.*, XHTML form) presented to the user, the latest metadata processing state results, previous metadata

[1] http://www.nokia.com/nokia/0,,2273,00.html

[2] http://www.symbian.com

[3] http://www.attws.com

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

[5] http://httpd.apache.org/

[6] http://jakarta.apache.org/tomcat/

annotated to the image during the current session, and the next possible set of states that the user might reach.

The session object has access to the metadata processing and retrieval module for generating metadata guesses. The particular set of metadata and retrieval algorithms chosen is based on the current annotation state of the user. The session object provides the algorithms with the information about the current state that can include the location, the previous annotations, the current user, and the current image.

Interaction between the session object and the user is carried out by the annotation UI module. The session object depends on the annotation UI module to generate the XHTML forms that are sent to the phone's browser. In this way, the session object provides the behavior and data retrieval functionality to the user, while the annotation UI provided graphical interaction functionality to the user. The session also has access to the repository of annotated images to store the metadata and images.

3.4.3 Annotation UI

The annotation UI module abstracts XHTML creation from the session object. It is aware of phone and browser specific issues like screen size and XHTML compatibility and optimization for the limited GPRS bandwidth. The annotation UI module also provides the functionality for parsing parameters from the HTTP request, capturing invalid values, and notifying the session of the inputted metadata.

3.4.4 Media and Metadata Processing and Retrieval

The metadata processing and retrieval module's main functionality is to provide the algorithm implementation for retrieving metadata using the current state of the annotation process and the repository of annotated images. The values returned by the metadata processing and retrieval module are the guesses sorted in order of highest probability. Depending on the request of the session object, the module provides two main sets of algorithms based on location information or creator information.

In general, the set of algorithms based on location information provide results based on the GSM network cellID at the time of image capture. The algorithms return a set of locations within that given cellID that have the highest probability of being correct. The location generation algorithms depend highly on the mapping of cellIDs to locations. These mappings consist of cellIDs mapped to a particular set of locations and locations mapped to a particular cellID.

The set of algorithms based on creator information associates weights to the creator of the image (*i.e.*, who took the image), the subject of the already taken images, and the frequencies of images already annotated by the user. The algorithms return a set of possible people that might be subject(s) of the image. Previous annotations of images and the subjects of those images were leveraged when generating the probabilities.

3.4.5 Storing of Annotated Images

The repository consists of an object-oriented database and a file system folder structure to store the image file and the metadata collected during annotation. The object-oriented database used is the Java open source Ozone 1.1 database[7]. All metadata including filename, cellID, time, user name, and annotations are stored in the object-oriented database.

When an annotation is created, the image file uploaded to the system is saved to a directory accessible by the web server and assigned to the user. The image file is given a unique file name once the successful upload is complete. The system-generated filename is stored in a facet object that is associated with any annotations for the image and with the image source. Access to the image source is available by generating a URL using the facet object and filename.

The Ozone database is mainly made up of facet classes, facet node classes, and facet object classes. The facet class is a container that represents one metadata facet, such as the people facet or location facet. The facet node class defines a node in the facet structure, which is tree data structure. Facet nodes have one parent and zero or more children. Each facet node also contains pointers to facet object class instances that have been annotated to that facet node. The facet object class represents anything that can be annotated to a facet. In our system, facet object class instances can represent images or users.

The object oriented database leverages the faceted metadata structures by storing annotations as objects that are annotated to the particular metadata facet node. By leveraging this structure we can quickly retrieve facet object class instances that are annotated to a particular facet node as well as retrieving the annotations (facet nodes) for one particular facet object class instance.

4. PROTOTYPE EVALUATION

The prototype infrastructure was initially designed, implemented, and iterated from March 2003 to August 2003 and underwent a four month trial with a group of 40 students in a required first year graduate course in Information Organization and Retrieval[8] in the fall semester of 2003 at the School of Information Management and Systems at the University of California at Berkeley. Each student was given a phone with the client-side program and class assignments to take pictures and annotate them using the prototype infrastructure. An additional 15 researchers used the MMM prototype as well during the trial.

The prototype system was evaluated on the basis of feedback and observations from the students using the phones. In addition to informal feedback, two initial user studies have been conducted. A qualitative study was made with 5 users of the MMM prototype. They were observed performing standard system tasks and were also interviewed after the tests. In addition to the user study, a study on the created metadata was made by reviewing the images and metadata created by all of the 40 students as well as logs of system usage and weekly self-reported usage behavior gathered through a web survey.

The following subsections describe in more detail the findings which arose from the studies.

4.1 System Evaluation

4.1.1 Integration of Components

Integrating the various components of the system brought forth issues involved in adopting the new phone technology. The

[7] http://www.ozone-db.org/

[8] http://www.sims.berkeley.edu/academics/courses/is202/f03/

Symbian operating system interfaces were not well documented and the software development kits for the platform were not well established. This caused a lot of time spent learning the technology and overcoming technical problems inherent in such a new platform. However, the prototype proved that the designed process could be implemented with current technology.

4.1.2 Network Unpredictability

The main usability problem with the prototype was the unpredictability and limited availability of the GPRS network used. Therefore, the network too often failed to transmit the image, metadata, or XHTML form data to the remote server. Users found this unpredictability troubling and at times struggled to gain a connection to the remote server. In combination with the limited bandwidth of the GPRS network, these local network problems were annoying to many users.

4.2 Usability Evaluation

The goal of the user studies was to provide feedback on the user interaction usability of the metadata creation process. The key issues we focused on were the user experience as the user interacts with the system using the *Image-Gallery* program and the XHTML browser. The other issue we focused on was the usability of the phone as an input device for image annotation.

4.2.1 Interaction Speed

Another main usability issue was also related to the GPRS network: once a user did have a network connection, the limited bandwidth became a bottleneck for the user interaction process due to the slow response time for image uploads and XHTML downloads.

Although the XHTML forms were optimized for the limited bandwidth, the user experience changed immensely from that initially conceptualized. We learned through testing that users often would set down their phones between image uploads or XHTML page loads, so as not to be bothered with staring at a busy, processing phone.

Our initial predictions were that the user interaction process would take 30-45 seconds. However, the tests indicated that the process was taking 3-5 minutes. This interfered too much with users' work flow and required their attention to the phone for too long of a duration.

Ultimately, many users found the slowness and unreliability of the network too distracting to their current activities to offer sufficient use value to them. Realizing this, one user in the study stated, "I would [have] used it more if it was as fast as my web browser [at home]."

Either a faster network or implementing more of the annotation functionality on the client side would make the user experience smoother and less vulnerable to network issues. Another approach would be to enable the user to concentrate on other tasks while the phone is communicating with the remote server, and thus lessen the frustration of waiting for a response.

4.2.2 Input Device Usability

The phone model itself created challenges for the usability of the prototype. As the only device for inputting both media and metadata, the basic usability of the phone was a major part of the overall user experience. Many users found text input on the phone's round keypad layout difficult and unfamiliar to use. Also, the central control button used for XHTML page navigation was awkward for some users.

The phone's XHTML browser does not cache images. Therefore, to minimize the use of the limited network bandwidth, the use of image thumbnails in the XHTML user interaction was disabled. The downside of turning off the image thumbnails was that without the visual reminder, the user often forgot the detailed subject matter that was to be annotated in the image.

The screen resolution (176x208 pixels) was also a challenge for the user interface. Especially in traversing the lists of choices for metadata the limited screen real estate was at times problematic. The list of metadata choices had to be carefully shortened without losing adequate information and intelligibility.

These issues are general usability problems of mobile phones: limited keyboard, relatively small screen size, and device dependent user interface conventions. In future work, the choice of implementing more functionality on the client-side would therefore mean more dependency on the phone models used and less portability to other models.

4.3 Metadata Evaluation

The objective of studying the generated metadata was to see what kind of leaf nodes users added to the metadata hierarchy and to what extent the system actually supports the annotation of semantic image content metadata.

4.3.1 Inconsistent Annotations

The challenge in allowing free text to be inserted into the metadata hierarchy is in the divergence of users' descriptive vocabularies. For example, one person may identify a stuffed monkey as a toy named "George". Another user may name it as a "Toy Monkey". Because of the limited screen space for listing all possible choices and the latency issues described above, users could not easily traverse down the metadata hierarchy to an exact level of detail where their description would ideally fit.

The result was that users input duplicate nodes with the same semantic meaning. This problem is seen more saliently when we consider a person's name. One person may know another by one name, say their given name, and annotate an image of that person with that name, while a different individual may know the same person by a nickname and annotate a photo of that person by that nickname. We then have two annotations for the same person. When a completely new user comes to use the system and knows the said individual by both names, which name are they to pick? It is therefore essential not merely to indicate annotations by name, but to indicate unique database objects which may have multiple nicknames that can be resolved to a single object.

Another metadata issue was identified in the user tests. Some users avoided adding new metadata, and instead, used the list of choices as prompts for allowable descriptions. When tested, users commented on the fact that to refrain from them having to use free text input, they would rather select something from the list. Therefore, they often used the list as a naming cue and searched the list for something that would closely fit the desired content in the image. This is very promising in that if users can reuse descriptions supplied by the system, greater semantic convergence and consistency can be achieved.

4.3.2 Generated Semantic Metadata

The purpose of the metadata creation process was to generate semantic content metadata about an image. Contrary to the hindrances the prototype had, it was still successful in providing images with metadata about their contents. Images in the repository had at least the automatically gathered metadata (*i.e.*, time, date, user name, and network cellID), and often at least one metadata facet annotated (*i.e.*, person, location, object, or activity). The photos our users took tended to be of locations, objects, or people. While the fact that only few images had more than one metadata facet annotated may likely have been due to slow network performance, based on a review of the captured photos and metadata, it may also be the case that the types of photos our users took could be well-described by using just one of the facets, *i.e.*, the main subject of the photo.

We also learned that a key motivation for users to annotate and share annotations is the ability to browse, search, and share media based on these annotations. The availability of this functionality was limited in the trial and available only on desktop-based web browsers in the final weeks. Future trials will include well-integrated metadata-based browsing, searching, and sharing functionality both on the phone and the web to support users in gaining direct benefit from their annotations.

5. FUTURE RESEARCH

Taking the user studies into account, further research on the annotation process will primarily focus on making the implementation prototype more usable. Special focus will be given to designing around the limited GPRS network bandwidth. The most promising solution seems to be implementing more of the user interaction on the client device rather than having majority of the dialog with a remote server over the network connection. However, this approach requires merging and synchronization of shared metadata information and will be less important as network speed and reliability improve.

Another area of future work is having more accurate location information available at the time of capture. The Bluetooth interface of phones enables the integration of a GPS device into the client. With more accurate location information the annotation process could be further automated, because the number of possible named locations a user would have to select from when annotating a location would be significantly smaller.

Also, future research issues include more sophisticated media and metadata similarity processing algorithms that would automate and simplify the annotation process further. We are continuing our development of algorithms that leverage similarities and patterns across spatial, temporal, and social contextual metadata to infer media content.

Another major area of research is the privacy issues involved in sharing images and metadata. How do the benefits of sharing information with everyone affect the need to keep some information private? How does the level of privacy affect the motivation to annotate?

Also, the applications using the metadata are directly related to the users' motivation for spending time annotating. Our user tests showed that users have little motivation to use the system for simply annotation, without any other reward to be obtained. However, presenting a small reward, such as being able to automatically view their annotated images on a website, was enough to bode enthusiastic responses. The design of applications and their effect on user motivation and the metadata ontology are major parts of future research. The applications currently under construction are a shared photo album, an image-based wish list, and mobile image games.

6. CONCLUSIONS

We have presented the design of a metadata creation process for images taken using a mobile camera phone. The design leveraged gathering of automatically available contextual metadata at the time of capture, it used metadata and media retrieval algorithms for reusing shared metadata and images, and it interacted with the user during image capture to confirm the system supplied metadata.

We built a prototype system to evaluate the designed metadata creation process. The MMM prototype was implemented using GSM phones that connected to a remote web server in the Internet via a GPRS network. The prototype was evaluated by observing 40 students and 15 researchers using the system and creating annotated images. A specific qualitative user study was conducted with 5 students as well as weekly usage logging and user surveys.

The evaluation showed that the creation process could be implemented with current mobile phone technology, and it facilitated the creation of semantic metadata at the time of image capture. While the limited bandwidth and unpredictability of the current GPRS network hindered the users' experience and the usability of the mobile phone model had some shortcomings, these obstacles are likely to be overcome in the next few years. What endures from our initial research is the proof of concept of a new approach, validated in a large scale user trial, of metadata creation, sharing, and reuse with camera phones to leverage the spatio-temporal and social context of media capture to infer media content and to do so using shared media and metadata resources and human-in-the-loop verification of system-supplied metadata. This approach promises to enable not only the solution of problems in consumer image management, but also the efflorescence of new applications and uses for annotated media in the mobile age.

7. ACKNOWLEDGMENTS

The authors would like to thank Amy Shuen, Katherine Chan, Preetam Murkherjee, Jaiwant Virk, Yuri Takhteyev, Ted Hong, Bruce Rinehart, Ray Larson, Olli Pitkänen, and Marko Turpeinen. This work is supported by British Telecom, AT&T Wireless, Futurice, and Nokia. This research is part of Garage Cinema Research (http://garage.sims/berkeley.edu) at the School of Information Management and Systems at the University of California at Berkeley and part of Mobile Content Communities research project at the Helsinki Institute for Information Technology (HIIT).

8. REFERENCES

[1] Aigrain, P., Zhang, H., and Petkovic, D. Content-based Representation and Retrieval of Visual Media: A State-of-the-Art Review. Multimedia Tools and Applications, vol. 3, 1996; pp. 178-202.

[2] Adobe. Photoshop Album. http://www.adobe.com/

[3] Bederson, B. PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and BubbleMaps. Proceedings of the 14th annual ACM symposium on User interface software and technology (UIST 2001), pp.71-80.

[4] Carson, C., Thomas, M., Belongie, S., Hellerstein, J.M., and Malik, J. Blobworld: A system for region-based image indexing and retrieval. In Proceedings of International Conference on Visual Information Systems, 1999.

[5] Chang, S-F. The Holy Grail of Content-Based Media Analysis. IEEE MultiMedia, 9 (2); pp. 6-10.

[6] Chen, G., and Kotz, D. A Survey of Context-Aware Mobile Computing Research. Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College, November, 2000.

[7] Davis, M. Editing Out Video Editing. IEEE MultiMedia, 10(2); pp. 54-64.

[8] Davis, M. Media Streams: An Iconic Visual Language for Video Representation. Readings in Human-Computer Interaction: Toward the Year 2000, R.M.Baecker et al., eds., 2nd ed., Morgan Kaufmann, 1995; pp. 854-866.

[9] Dorai, C. and Venkatesh, S. Computational Media Aesthetics: Finding Meaning Beautiful. IEEE Multimedia, 8 (4). pp. 10-12.

[10] Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. The Vocabulary Problem in Human-System Communication. Communications of the ACM, 30 (11). pp. 964-971.

[11] Girgensohn, A., Adcock, J., Cooper, M., Foote, J., and Wilcox, L. Simplifying the Management of Large Photo Collections. Human-Computer Interaction INTERACT'03, IOS Press, 2003; pp.196-203.

[12] Kuchinsky, A., Pering, C., Creech, M.L., Freeze, D., Serra, B., and Gwizdka, J. FotoFile: A Consumer Multimedia Organization and Retrieval System. In Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 99). (Pittsburgh, Pennsylvania, 1999), ACM Press; pp. 496-503.

[13] Naaman, M., Paepcke, A., and Garcia-Molina, H. From Where to What: Metadata Sharing for Digital Photographs with Geographic Coordinates. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA and ODBASE. Springer-Verlag 2003. pp.196-217.

[14] Nokia Image Upload Server API, version 1.1, available at http://www.forum.nokia.com/

[15] Schilit, B., Adams, N., and Want, R. Context-aware computing applications. In Proceedings of IEEE Workshop on Mobile Computing Systems and Applications, Santa Cruz, California, December 1994. IEEE Computer Society Press; pp. 85-90.

[16] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-Based Image Retrieval at the End of the Early Years, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, 2000; pp. 1349-1380.

[17] Stentiford, F. W. M. An attention based similarity measure with application to content based information retrieval. In Storage and Retrieval for Media Databases 2003, M. M. Yeung, R. W. Lienhart, C-S Li, Editors, Proc. SPIE Vol. 5021, pp. 20-24.

[18] Toyama, K., Logan, R., Roseway, A., Anandan, P. Geographic Location Tags on Digital Images, Proceedings of the 11th ACM International Conference on Multimedia, ACM Press; pp.156-166.

[19] Vartiainen, P. Using Metadata and Context Information in Sharing Personal Content of Mobile Users. Master's Thesis, Department of Computer Science, University of Helsinki, Finland, 2003.

[20] Wenyin, L., Dumais, S.T., Sun, Y.F., Zhang, H.J., Czerwinski, M.P., Field, B. Semi-automatic image annotation. In Proceedings of Interact 2001, Eighth IFIP TC.13 Conference on Human Computer Interaction. July 2001.

[21] XHTML Mobile Profile Specification, version 29-Oct-2001, available at http://www.wapforum.org/

[22] Yee, K-P., Swearingen, K., Li, K., and Hearst, M. Faceted Metadata for Image Search and Browsing, Proceedings of the Conference on Human factors in computing systems, ACM Press; pp. 401-408.

Appendix 1. Sequence Diagram of the Annotation Process