From Context to Content: Leveraging Context for Mobile Media Metadata

Marc Davis[1], Nathan Good[1], Risto Sarvas[2]

[1]*University of California at Berkeley*
School of Information Management and Systems
Garage Cinema Research
102 South Hall, Berkeley, CA 94720-4600
{marc, ngood}@sims.berkeley.edu

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

Abstract

The recent popularity of mobile camera phones allows for new opportunities to gather important metadata at the point of capture. This paper describes and demonstrates a method for generating metadata for images using spatial, temporal, and social context. We describe a system we implemented for inferring location information for pictures taken with camera phones. We propose that leveraging contextual metadata at the point of capture can bridge the problems of the semantic and sensory gaps. In particular, combining and sharing spatial, temporal, and social contextual metadata from a given user and across users allows us to make inferences about media content .

1. Introduction

Past work in content-based image retrieval has focused on databases, computer vision, and information retrieval [1]. As the world evolves into a more distributed, networked system full of wireless connected media capture devices, we are now able to sense, infer, and learn the context of creation and use of media. Access to and processing of contextual information available at the point of capture allows us far greater leverage in solving a fundamental problem of content-based image retrieval: the *semantic gap* between the low-level features that can be automatically parsed from media signals and the semantically meaningful descriptions that users want to use when searching for and managing media [1, 2].

Mobile devices are designed to take into account the users' physical environment and usage situations and can ultimately enable us to infer media *content* from the *context* of media creation and use. Camera phones can also leverage location services to supply spatial metadata for mobile imaging (e.g., GSM network cellID, EOTD-based location, GPS, etc.). Furthermore, by utilizing the networking, interaction, and communications capabilities of camera phones, collaborative, cooperative image annotation applications are possible between the system and the user and among users.

We have developed a camera phone image annotation system that offers unique opportunities for media semantics by enabling annotation at the time of image capture, adding some contextual metadata automatically, leveraging networked metadata resources, and enabling iterative metadata refinement on the mobile media device [3, 4]. A fundamental part of this system is an inference engine that leverages the spatial, temporal, and social context of media creation and use to infer metadata about media content.

To illustrate this approach it is helpful to use an example. Imagine that a father is taking a picture of his three children, and we would like to add this information to the image for future use and retrieval. Using the spatial context of capture, we can tell that this picture is being taken at the father's house. From a profile of past images, we are able to infer that there is a strong probability that pictures in the house include his children, and are limited to a very small subset of other possible people. From the temporal context, we can tell that it is a Saturday, a day when the father and the children are likely to be home together further constraining the space of likely options. From the social context, we can determine that if the father is taking the pictures, it is most likely of his children. With all of

these values included in the inference engine, we are able to make a reasonable guess that the picture is indeed of the father's children, and then interactively verify this information with the user. While we are not 100% sure that the father and children are together and that he is taking photographs of them, from past photos taken in similar spatial, temporal, and social contexts we can make a good guess about the content of the image (e.g., who is in the photo), especially if the user indicates that he has taken a shot of people. In the case tht our system guesses incorrectly about the image content, the user has the option to correct this, and then we add this to our profile of the person to help improve future "context-to-content" inferencing.

The power of this approach is that it allows us to closely align a user's interpretation of an image with the actual content of the image, something that current algorithms cannot do. In addition, we demonstrate an architecture that allows us to do this at the point of capture. In this paper, we discuss related work in contextual metadata and describe a system we built connecting 55 Nokia 3650 camera phones and a metadata server that infers media content semantics from spatial, temporal, social contextual metadata.

2. Related Work

In Smeulders et al.'s survey of content-based image retrieval [1], the "semantic gap" and "sensory gap" describe two major obstacles image retrieval systems still must overcome in order to gain widespread acceptance. The sensory gap is described as the gap between an object and the computer's ability to sense and describe that object. For example, for some computational systems a "car" ceases to be a "car" if there is a tree in front of it, effectively dividing the car in two from the machine's perspective. In addition to problems with object occlusions, signal-based parsing of image databases cannot easily differentiate perceptually similar images that are in fact of different objects or unify perceptually dissimilar images which are in fact views of the same object. As we shall see below, it is contextual knowledge which enables computational systems to bridge the sensory gap.

The semantic gap is described as the gap between the high-level semantic descriptions humans ascribe to images and the low-level features that machines can automatically parse [2]. For example, a picture of a man tossing a red ball to a dog would be "seen" by a vision system as a series of moving color regions. The relationship between the man, the dog, the location where the ball is being thrown and the significance of this event to the person taking the picture are all gone.

As described by [1], content-based image retrieval has attempted to work around these problems using a variety of methods. For the sensory gap, domain and world knowledge are explicitly built into the system. Knowledge that describes physical laws, laws about how objects behave and how people perceive them, and other supporting rules and categories are incorporated into the system in the hope of improving recognizers and helping machines bridge the sensory gap. To date this type of knowledge-based approach has only really been viable for highly constrained, controlled, and regularized domains such as industrial automation applications. With the semantic gap, the most common means of attempting to solve the problem are by adding captions or annotations to images. This however, is a costly and tedious process that requires many hours of effort, tweaking of machine algorithms, and careful watch over vocabulary and content to make sure that the images are tagged correctly. In addition, most previous work in image annotation is done long after the image has been created, where it is most difficult to extract useful information about the image.

Recent work has looked at addressing parts of these problems by incorporating additional metadata with the image, most noticeably spatial context. Toyama et al.'s research [5] enables users to tag their photos with GPS data and share these spatially indexed images with others across the world via a web site. Combined with a map, the system allows users to effectively view other people's images from locations they know of or are interested in. Recent work at Stanford [6] allows devices to share location information and labels for photographic images. Like our own work [3, 4], it uses location to determine what labels other photographs taken in a similar location should have.

In ubiquitous computing research, researchers have attempted to use location information to infer context as well as activities of people operating inside of their environments. Research by Dey [7] describes how to infer users' actions by the context of their locations, and possibly by looking at patterns of what they have done previously. In addition, related work has looked into using inference engines to infer location based on a system of rules and constraints [8]. Unlike this prior work in context-aware and ubiquitous computing, our research aims to utilize context-aware computing to solve long standing problems in media asset management. By focusing on the context of media creation using mobile devices, we can use insights from context-aware computing about how to capture and

model context (especially where, when, and who) to solve once intractable problems in media content analysis, retrieval, sharing, and reuse.

We developed our Mobile Media Metadata (MMM) system [3, 4] at the same time as [6, 9], but in contradistinction to [6], we use a faceted semantic ontology for media description, and moving beyond [6, 9] leverage social, temporal, and spatial contextual metadata to make inferences about media content. In [3] we provide a comprehensive overview of our MMM prototype; in [10] we describe an evaluation of the users experience with the MMM prototype; and in this paper we describe MMM's "context-to-content" inferencing system.

3. Bridging the Semantic and Sensory Gaps through Contextual Metadata

In content-based image retrieval, most attempts at bridging the semantic and sensory gaps have focused on deriving media semantics after the media has been produced (i.e., created and edited) [2]. We have explored bridging the semantic gap by leveraging the point of media capture in our research on "Active Capture" [10]. With the advent of mobile phones with cameras, we have a new opportunity to capture and infer media semantics at the time the image is captured. In leveraging the *context* of media creation to help bridge the semantic and sensory gaps, we can take advantage of three aspects of image context that seem to have special salience in most consumer photos: when, where, and who. By choosing temporal, spatial, and social context, we were able to use the existing camera phone and network infrastructure to gather this information, and incorporate user interaction to adjust and add more information when needed.

4. System Description

We created a prototype "Mobile Media Metadata" (MMM) system that allows users to annotate pictures on Nokia 3650 camera phones. MMM has been deployed since September 2003 and was used by 40 graduate students and 15 researchers at the University of California at Berkeley's School of information Management and Systems in a required graduate course entitled "Information Organization and Retrieval" co-taught by Prof. Marc Davis and Prof. Ray Larson. Students used the MMM prototype and developed personas, scenarios, storyboards, metadata frameworks, and presentations for their concepts for mobile media and metadata creation, sharing and reuse applications (www.sims.berkeley.edu/academics/courses/is202/f03/phone_project/index.html).

The users were asked to take pictures and annotate these pictures using a simple semantic ontology so others could view and reuse their metadata. From experience, we knew that the process of annotating images was tedious and error prone, so we wanted to design a system that would provide users an easier way to annotate them. As depicted in Figure 1, MMM gathers metadata from the context of capture, suggests additional metadata based on the repository of similar annotated images, and then interacts with the camera phone user to confirm, reject, or augment the system-supplied metadata.

We saved all of the students' data and metadata to a single database to facilitate sharing and correlation of

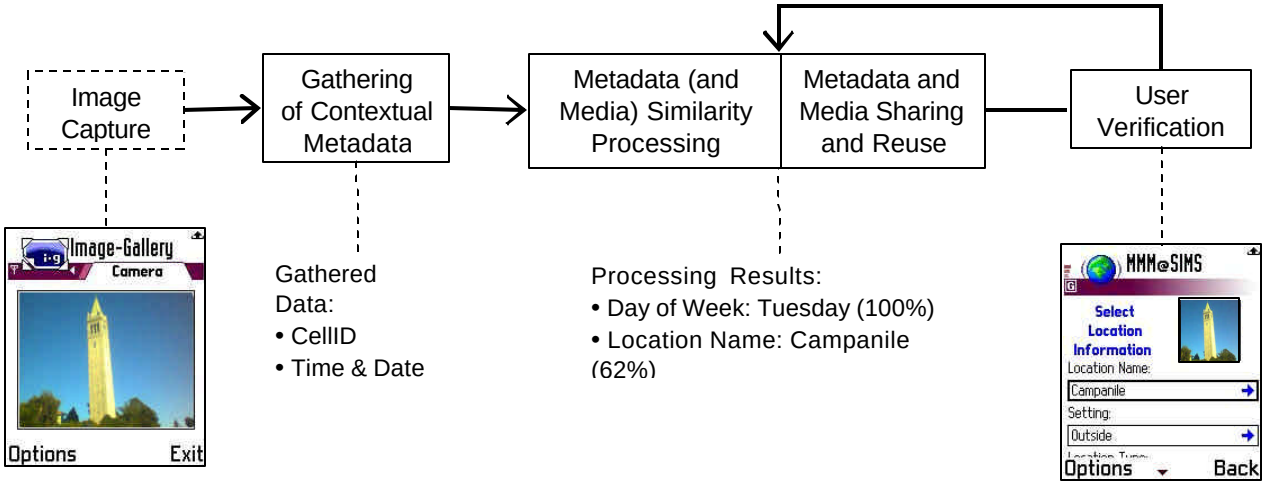

Figure 1. Mobile Media Metadata image annotation process

information. For example, if the majority of users have been standing at a spot and they all took pictures of the Campanile at UC Berkeley, there is a strong chance that if the another user is standing in the same spot or somewhere nearby, then they are also taking a picture of the Campanile

By exploiting regularities in spatial, temporal, and social contexts shared by a network of camera phone users, we were able to solve the problems characterized by the sensory and semantic gaps. Today it is impossible for signal-based analysis alone to be able to tell that an off-white, vertically-oriented box of pixels in an image is the Campanile at UC Berkeley, especially if it is taken from multiple angles, or on different days with different weather and lighting conditions. Furthermore, if an image analysis algorithm was given similar looking photos of three towers from different geographic locations, it wouldn't know if they were of the same tower or not. By using the spatial context of where the image is taken, we are able to infer that different images taken in the vicinity of the Campanile are of the Campanile and know that they are not of, for example, the Washington Monument.

It is important to note that we do not know where the user is pointing the camera and are not using image processing to determine if the image exists. Rather, we are relying on the most probable content of the image based on prior history and knowledge of what has been captured at that point. For this reason, even a picture that is taken far away from the actual location can be associated with the depicted object and location. For example, if a picture is taken from a vantage point miles away from the campus that that has a good view of the Campanile, from a user's perspective the image is of the Campanile, so the location of the image content should reflect that and not the "vantage point on a mountain". This distinction between the "camera location" (where the photo is taken from) and the "subject location" (the location of the subject of the photo) is a key differentiation for context-aware media systems and applications.

MMM combines a GSM/GPRS camera phone and a remote web server in a client-server architecture (See Figure 2). Using our client software on the phone, the user captures the image and selects the main subject of the image (*Person*, *Location*, *Activity*, *Object*) before uploading it to the server. The server receives the uploaded image and the metadata gathered at the time of capture (*main subject*, *time*, *date*, network *cellID*, and *user name*). Based on this metadata, the server searches a repository of previously captured images and their respective metadata for similarities. The images and metadata in the repository are not limited to the user's own images and metadata, but contain every user's annotated media to leverage the advantages of shared metadata. For example, the probability that a person is present to be photographed at a given place and time (e.g., home on the weekend) may be raised by a photographer other than the MMM user having photographed that person in that place before around that time.

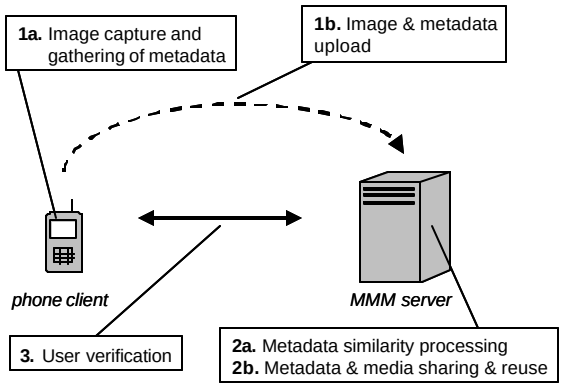

Figure 2. Mobile Media Metadata (MMM) system overview

Using the information from previously captured images that had similar metadata, the server program generates educated guesses for the user (i.e., selection lists with the most probable metadata first). The user receives the server-generated guesses for verification, and selects the correct metadata. As the user and system continue through the verification process, the verified metadata is sent back to the server for further processing such that the server can take advantage of the verified metadata to generate better guesses as the process progresses. For example, my now verfied metadata about location will likely help the system provide a much better guess about who may be in the photo. The interaction between the user and the server ends when either the user terminates it or the server does not generate any more guesses. Below we describe the system implementation in more detail by dividing it into the main parts of the metadata creation process.

4.1. Image capture and metadata gathering

The client side image capturing, user selection of main subject, automatic gathering of metadata, and communication with the server were implemented in a C++ application named *Image-Gallery*. It was developed in cooperation with Futurice (www.futurice.fi) for

the Symbian 6.1 operating system on the Nokia 3650 phone. The user captures the image using the *Image-Gallery* which then automatically stores the *time, date*, GSM network *cellID*, and the *user name*. The image and metadata upload process was implemented in *Image-Gallery* and on the server side using the Nokia Image Upload API 1.1.

4.2. Metadata similarity processing

The server side metadata similarity processing was implemented in a Java module that provides a set of algorithms for retrieving metadata using the metadata of the image at hand and the repository of previously annotated images. The values returned by the metadata processing and retrieval are the guesses sorted in order of highest probability. In the MMM system we implemented two main sets of algorithms: location guessing based on spatio-temporal patterns, and person guessing based on social patterns using the mobile phone user name as a person identifier. Spatial, temporal, and social similarity intersect in myriad ways as illustrated in Figure 3.

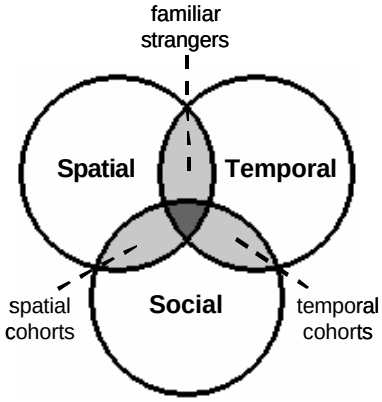

Figure 3. Connecting spatial, temporal, and social metadata

The patterns in where, when, and with and of whom individuals, social groups, and cohorts take photographs have discernible regularities that we use to make inferences about photo content. For example, based on regularities in contextual and user supplied metadata, the system should predict that it is far more likely that a parent would be taking a photo of one of their young children at home on the weekend vs. at work during the week. These patterns influence the rank order of suggested persons who may have been photographed at a given place and time.

MMM uses a weighted combination of spatial, temporal, and social metadata to infer location and person. In its current implementation, it uses a very simple linear combination of several variables that we determined through past experience would be good predictors of image content based on contextual metadata. We chose the simplest implementation for speed and to determine if the variables that we had chosen were useful predictors. If they worked in the most basic case, then we could develop more sophisticated implementations that combined reasoning engines and machine-learning algorithms.

4.2.1. Inferring Spatial Context

It is important to note the distinction between our goals and existing methods used to determine location. We were not only trying to infer the location where the image was taken (i.e., camera location), but also we were trying to infer the location of what the image was taken of (i.e., image content location). Leveraging regularities in a given user's and in a community of users' photo taking behaviors helps us address this challenge. Figure 1 shows an example of how our system attempts to infer the location of image content.

We chose weights for our location-determining features based on our past experience and intuition, and then adjusted them through a process of trial and error. For example, it seems intuitive that if two pictures are being taken in the same location within a certain time frame (e.g., a few minutes for pedestrian users), they are probably in or around the same location. Another factor we considered is the intersection of spatial, temporal, and social metadata in determining the location of image content. Within a given cellID, patterns of being in certain locations at certain times with certain people will help determine the probability of which building in an area I might be in and/or photograph, if it is, for example, my place of work. In future iterations we hope to use rule-based constraint and inference engines to aid reasoning, and machine learning algorithms to learn from past performance to optimize and adjust the relative importance of the various location-determining features.

4.2.2. Inferring Social Context

In the example above we described how a system could infer that a father is taking a picture of his children by reasoning about who he normally takes pictures of at a given location. In this case, the system knows something about the father's social network

based on explicit user-created representations (e.g., "son", "daughter", etc.) and metadata about who was in photos taken by whom, and how this network of relations intersect with particular places and times.. We developed a simple system of determining who was likely to be in a picture by using a weighted combination of various factors that we felt were important and useful based on experience and intuition. We decided to use three main relationships in determining whether a person was present or not, and then use these to create other relationships that may be used to reason about people as well. We looked at whether a person A had ever taken a picture of person B, whether a person B had ever taken a picture of person A, and whether A and B have been in the same picture. From these initial relationships, we could also look at whether they have taken pictures of each other at the same time, within a period of time, at the same places, in different places, as well as several others. Also, we looked at the time in between pictures to determine how likely it was that the person in the last picture was still around for the next.

In future work, we plan to perform an experimental validation of our techniques, as well as comparison tests for various types of inference engines and learning approaches to come up with interesting approaches for dealing with predicting image content from the context of capture and use.

4.3. Metadata and media sharing and reuse

One of the main design principles in the MMM system 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. Therefore, the metadata processing module can reuse the metadata in the system in generating the guesses for the user.

While shared metadata is a useful concept, it is important to recognize the privacy concerns around sharing user profile information with others. Data such as time, place, and location all can potentially violate people's privacy. While in our current system privacy hasn't been an issue, we recognize that at a larger scale privacy will be central to the systems we are trying to build. We hope to alleviate many concerns by aggregating data whenever possible, and intend on exploring additional means for preserving privacy in future work.

The images and their respective metadata are stored in an open source object-oriented database (Ozone 1.1) on the server. The metadata is stored in a faceted hierarchical structure. The objective of the faceted structure is for the facets to be as independent of each other as possible, in other words, one facet can be described without affecting the others. In our structure the facets were the main subjects of the image: *Person*, *Location*, *Object*, and *Activity*.

4.4. User verification

The user verification and system responses were implemented in XHTML forms. After uploading the image and metadata, the client-side *Image-Gallery* program launches the phone's XHTML browser to a URL given by the server during the uploading. After the server creates the metadata guesses to facilitate the user's annotation work, it creates XHTML pages from the guesses for the client-side browser to present to the user. The dialog between the server and the user is then implemented in the form data sent from the phone to the server, and the XHTML pages created by the server that are rendered by the phone's browser.

4.5. Bootstrapping the system

As with any inferencing system, it is important to be able to provide value with even sparse datasets by bootstrapping it with known values. Temporal, spatial and social context can be bootstrapped prior to computation. The relative frequencies, clusters, and patterns of times in which a user's prior photos have taken can be automatically determined from JPEG file headers and used to bootstrap the inferencing system. POI (points of interest) databases and any existing geo-coded image collections [5] data can be used to narrow down the choices of a given location. In addition, popular POIs can be weighted more heavily in the beginning to assist inferencing with sparse datasets. Social context can be similarly bootstrapped by the system initially asking people who they most take pictures of, or by determining their photo-social relations through other means such as data from a social network service such as Friendster or by harvesting names from a user's already annotated images. In future work we plan to evaluate the effectiveness of approaches to bootstrapping metadata.

5. Conclusion

Using spatial, temporal and social context, we infer media content during image capture, enabling our system to bridge the semantic and sensory gaps. With more sophisticated inference algorithms, we will add even more metadata at capture time to create reusable and searchable media components at the beginning of the media production cycle.

In future work we will be evaluating the effectiveness of our algorithms and assumptions for context-to-content inferencing. In addition, we hope to develop and refine means of bootstrapping the system to minimize human intervention whenever possible, yet maximize its effectiveness.

6. Acknowledgements

The authors would like to thank British Telecom, AT&T Wireless, Futurice, and Nokia for their support of this research and the members of Garage Cinema Research at the UC Berkeley School of Information Management and Systems.

7. References

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

[2] C. Dorai and S. Venkatesh, "Computational Media Aesthetics: Finding Meaning Beautiful," *IEEE Multimedia*, vol. 8, pp. 10-12, 2001.

[3] R. Sarvas, E. Herrarte, A. Wilhelm, and M. Davis, "Metadata Creation System for Mobile Images," to be presented at MobiSys2004, Boston, MA, Forthcoming 2004.

[4] A. Wilhelm, Y. Takhteyev, R. Sarvas, N. Van House, and M. Davis. "Photo Annotation on a Camera Phone." In: *Extended Abstracts of the 2004 Conference on Human Factors in Computing Systems (CHI 2004) in Vienna, Austria.* ACM Press, 1403-1406, 2004.

[5] K. Toyama, R. Logan, & A. Roseway, "Geographic Location Tags on Digital Images," MM2003, Berkeley, CA, 2003.

[6] M. Naaman, A. Paepcke, and H. Garcia-Molina, "From Where to What: Metadata Sharing for Digital Photographs with Geographic Coordinates," CoopIS 2003, Catania, Sicily.

[7] A. K. Dey, "Understanding and Using Context," *Personal and Ubiquitous Computing Journal*, vol. 5, pp. 4-7, 2001.

[8] R. Hull, B. Kumar, D. Lieuwen, P. F. Patel-Schneider, A. Sahuguet, S. Varadarajan, and A. Vyas, "Enabling Context-Aware and Privacy-Conscious User Data Sharing," MDM 2004, Berkeley, CA, 2004.

[9] P. Vartiainen, "Using Metadata and Context Information in Sharing Personal Content of Mobile Users," Master's Thesis, University of Helsinki, Finland, 2003, pp. 67.

[10] M. Davis, "Active Capture: Integrating Human-Computer Interaction and Computer Vision/Audition to Automate Media Capture," ICME2003, Baltimore, MD, 2003.