Mobile Media Metadata: Metadata Creation System for Mobile Images

Marc Davis
University of California at Berkeley School of Information Management and Systems Garage Cinema Research
314 South Hall, Berkeley, CA, USA 94720-4600
+1 510 643-2253 http://garage.sims.berkeley.edu/
marc@sims.berkeley.edu

ABSTRACT
In the 2003, more camera phones were sold worldwide than digital cameras. With this new platform, we can leverage regularities in the spatio-temporal context and social community of media capture and use to infer media content. We created and deployed a "Mobile Media Metadata" (MMM) prototype on Nokia 3650 camera phones with 55 users that uses "context-to-content" inferencing, a shared metadata ontology, and user interaction at the point of capture to effectively infer media content annotations, specifically the semantic description of the location of the subject of users' photos.

Categories and Subject Descriptors
H.5.1 [**Information Interfaces and Presentation (*e.g.*, HCI)**]: Multimedia Information Systems; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.1.2 [**User/Machine Systems**]: Human Factors; H.5.2 [**Information Interfaces and Presentation (*e.g.*, HCI)**]: User Interfaces; H.5.3 [**Information Interfaces and Presentation (*e.g.*, HCI)**]: Group and Organization Interfaces; I.4.m [**Image Processing and Computer Vision**]: Miscellaneous.

General Terms
Algorithms, Design, Human Factors.

Keywords
Mobile Camera Phones, Contextual Metadata, Content-Based Image Retrieval, Context-to-Content Inference, Wireless Multimedia Applications, Location-Based Services

1. INTRODUCTION
In 2003, more camera phones were sold worldwide than digital cameras. The advent of this new platform for multimedia computing enables us to explore new approaches to solving long standing problems in media asset management [4]. Camera phones combine: media capture (images, video, audio); programmable processing using standard operating systems, programming languages, and APIs; wireless networking; rich user interaction modalities; time, location, and user contextual metadata; and personal information management functions. They enable us *at the point of media capture* to automatically gather contextual metadata (time, location, and user) and to interact with the user to confirm and augment additional system-inferred metadata.

As people go through the world capturing media, they effectively make paths in *space* (where they are and the location of what they photograph), *time* (when they are in locations and when they take photos), and *social space* (by, with, and of whom photos are taken). As conceptually illustrated in Figure 1, these paths and intersections in space, time, and social space have statistical regularities that we can use to make inferences about photo content, effectively leveraging the overlaps, clusters, and patterns in individual and group phototaking behavior. For example, in certain locations, such as tourist destinations like UC Berkeley, visitors are highly likely to photograph the Campanile Tower; or when a parent is at home on the weekend they are more likely to photograph their children than a co-worker from the office.

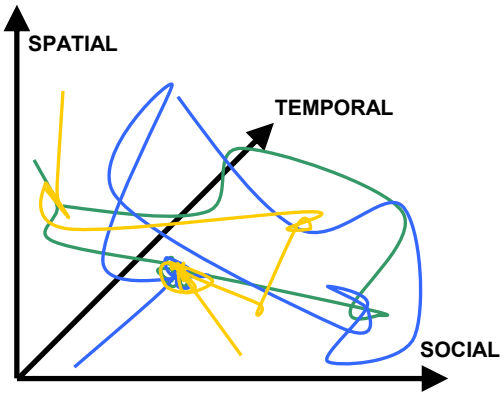

Figure 1. Conceptual Diagram of Paths and Intersections in Spatio-Temporal-Social Space

In the Mobile Media Metadata (MMM) project [1, 2, 3, 5] we built a system inspired by these insights and the possibilities of the new camera phone platform to effectively answer the question of "what did I just take a picture of?" by collecting and analyzing both automatically gathered contextual metadata and user-verified annotations from a shared ontology. In MMM we:

- Gather all automatically available information at the point of capture (time, location info expressed as CellID, phone user)
- Use metadata similarity algorithms to find similar media that has been annotated before
- Take advantage of this previously annotated media to make inferences about the content of the newly captured media
- Interact with the phone user at the point of capture to confirm and augment system-inferred metadata

Even if just a few users initially enter or verify metadata, this metadata can be propagated through spatio-temporal-social

MM'04, October 10–16, 2004, New York, New York, USA.

inferencing to similar photos taken by other users thus leveraging the network effects of shared context and metadata.

2. SYSTEM OVERVIEW

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. The MMM system connects Nokia Series 60 GSM/GPRS camera phones and a remote web server in a client-server architecture (See Figure 2) to enable context-to-content inferencing and annotation at the point of image capture. Using our client software on the phone, the user captures a photo and immediately selects the main subject of the photo (*Person*, *Location*, *Activity*, *Object*) before uploading it to the server. The server receives the uploaded photo and the metadata gathered at the time of capture (*time*, *date*, *CellID*, and *username*). Based on this metadata, a server-side metadata similarity algorithm compares the uploaded photo's metadata to a database of previously captured photos and their respective metadata to infer the likely metadata for the new photo. The photos and metadata in the database are not limited to the user's own, but contain all users' annotated media to leverage shared metadata. Using information from previously captured photos that have similar contextual metadata, the server makes inferences about additional metadata which it presents to the user on the phone's XHTML browser in selection lists sorted by probability. The user then verifies the server-generated guesses by selecting or augmenting the system-supplied metadata.

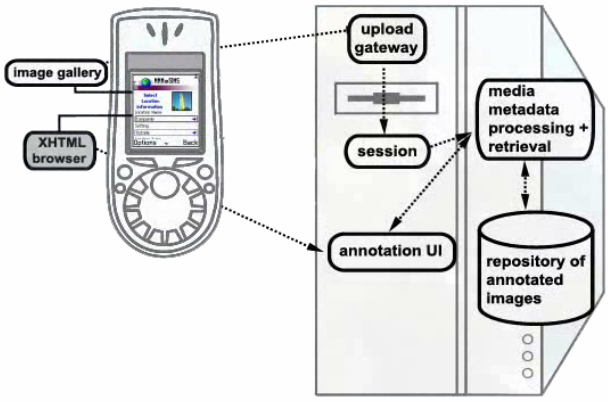

Figure 2. Mobile Media Metadata System Architecture

MMM's metadata is stored in a facetted semantic ontology. The top-level facets are the possible main subjects of the photo: *Person*, *Location*, *Object*, and *Activity*. We bootstrapped MMM's ontology by prepopulating it with a number of POIs from the Berkeley campus and the Bay Area, the names of the registered users of the system, and a small set of high-level object and activity descriptors. We also allowed users to add new terms to the common ontology, thus enabling shared bootstrapping.

We also implemented and tested a location inferencing algorithm that would make guesses about the named location of the subject of a user's photo within a given CellID. MMM's location guesser generates a sorted list of likely locations based on the output of several subalgorithms. Each subalgorithm generates a probability for each location associated with the user's current CellID. The probabilities are multiplied by a weight associated with each subalgorithm and added together. The resulting list of locations is then sorted by probability score. Currently guesses are based on the output of six subalgorithms: *Same User* (locations that have been photographed by the same user and in which the user has been photographed); *Delta(Time)* (how recently this location has been photographed by the user); *Same Time of Day* (how frequently this location was photographed at the same time of day across all users); *Same Date* (how frequently this location was photographed on the same day of the month across all users); *Same Day of Week* (how frequently this location was photographed on the same day of the week across all users); and *Same Day Class* (scores based on same type of day: "weekend" or "weekday"). Our location guesser was able to guess the correct location for a photo's subject very effectively. Exempting the times when a user first enters a new location into the system, MMM guessed the correct location of the subject of the photo (out of an average of 36.8 possible locations) 100% of the time within the first four guesses, 96% of the time within the first three guesses, 88% of the time within the first two guesses, and 69% of the time as the first guess.

3. CONCLUSION

In the Mobile Media Metadata system we implemented and evaluated a new approach to inferring media content from the spatial, temporal, and social context of media capture. We integrated media capture and analysis at the point of media creation; leveraged spatial, temporal, and social contextual metadata across individual users and groups to infer media content; and supported user-system interaction at the point of capture to enable "human-in-the-loop" algorithm design.

4. ACKNOWLEDGMENTS

The author would like to thank British Telecom, AT&T Wireless, Nokia, Futurice, and the Helsinki Institute for Information Technology for their support of this research, the Garage Cinema Research Mobile Media Metadata project team, and the other members of the MMM Video Team: William Tran, Arian Saleh, Erick Herrarte, Anita Wilhelm, Ali Sant, Dan Perkel, and Nick Reid.

5. REFERENCES

[1] Davis, M., King, S., Good, N, and Sarvas, R. From Context to Content: Leveraging Context to Infer Media Metadata. In *Proc. of ACM MM 2004* (New York, NY, October 10-16, 2004). ACM Press, New York, NY, Forthcoming 2004.

[2] Davis, M. and Sarvas, R. Mobile Media Metadata for Mobile Imaging. In *Proc. of ICME 2004* (Taipei, Taiwan, June 27-30, 2004). IEEE Press, New York, NY, 2004.

[3] Sarvas, R., Herrarte, E., Wilhelm, A., and Davis, M. Metadata Creation System for Mobile Images. In *Proc. of MobiSYS 2004* (Boston, MA, June 6-9, 2004). ACM Press, New York, NY, 2004, 36-48.

[4] 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 Trans. on Pattern Analysis and Machine Intelligence, 22, 12* (Dec. 2000), 1349-1380.

[5] Wilhelm, A., Takhteyev, Y., Sarvas, R., Van House, N., and Davis, M. Photo Annotation on a Camera Phone. In *Extended Abstracts of CHI 2004* (Vienna, Austria, April 24-29, 2004). ACM Press, New York, NY, 2004, 1403-1406.