# From Context to Content: Leveraging Context to Infer Media Metadata

Bibliographic Reference:

Marc Davis, Simon King, Nathan Good, and Risto Sarvas. "From Context to Content: Leveraging Context to Infer Media Metadata." In: *Proceedings of 12th Annual ACM International Conference on Multimedia (MM 2004) Brave New Topics Session on "From Context to Content: Leveraging Contextual Metadata to Infer Multimedia Content" in New York, New York*, ACM Press, 188-195, 2004.

# From Context to Content: Leveraging Context to Infer Media Metadata

Marc Davis[1], Simon King[1], Nathan Good[1], and Risto Sarvas[2]

[1]University of California at Berkeley
School of Information Management and Systems
314 South Hall, Berkeley, CA, USA 94720-4600
+1 510 643-2253
http://garage.sims.berkeley.edu/
{marc, simonpk, ngood}@sims.berkeley.edu

[2]Helsinki Institute for Information Technology (HIIT)
P.O. Box 9800
02015 HUT, Finland
+358 9 694 9768
http://www.hiit.fi/risto.sarvas/
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 a method for generating metadata for photos using spatial, temporal, and social context. We describe a system we implemented for inferring location information for pictures taken with camera phones and its performance evaluation. We propose that leveraging contextual metadata at the point of capture can address 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.

## Categories and Subject Descriptors
H.1.2 [**User/Machine Systems**]: Human Factors; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.1 [**Information Interfaces and Presentation (*e.g.*, HCI)**]: Multimedia Information Systems; 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
Multimedia researchers have been trying to solve the problems of content-based image retrieval and media asset management for well over the past decade [1, 14]. It is time to acknowledge that this research has not delivered on its promises. Neither fully automatic signal-based analysis nor manual annotation of media content has provided a workable solution to content-based multimedia access. A new direction and new solutions are needed. In order for media to be as accessible as text, descriptions of its content and structure (i.e., metadata) must be created that are computationally and humanly usable. Unfortunately, the low-level features that current algorithms can extract are not sufficient to meet the needs of how humans want to search for and use media content. This "semantic gap" is endemic to current multimedia information systems [8].

Most prior research in multimedia content analysis has suffered from assumptions that have impeded progress in the field and in making systems that can address users' needs. These assumptions are: 1) media capture and media analysis are disconnected in time and space such that media must be analyzed long after they have been captured (and therefore effectively removed from their context of creation and the users who created them); 2) contextual metadata about the capture and use of media are not available to media analysis and hence all analysis of media content must be focused on the media signal; and 3) multimedia content analysis must be fully automatic and avoid user involvement (and therefore miss out on the possibility of "human-in-the-loop" approaches to algorithm design). These assumptions can be overcome by shifting the paradigm of media processing from an a-contextual, fully automatic model to a model that leverages the spatio-temporal-social context of media creation and use as well as interaction among devices and people (especially by taking advantage of new programmable networked mobile media capture devices). By making this shift in method and focus that takes advantage of new technology platforms for media creation, we have the promise of solving long-standing challenges in multimedia systems.

The devices and usage contexts of media capture are undergoing rapid transformation from the traditional camera-to-desktop-to-network pipeline to an integrated mobile media experience. We now see a new class of networked media capture devices (typified by camera phones) that combines: 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 metadata; and personal information management functions. The confluence of this functionality at the point of media capture means that we have a unique opportunity to attempt to solve once intractable


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

problems in media asset management and multimedia information retrieval in an entirely new way. We can now leverage the spatio-temporal *context* and social *community* of media capture and use to infer media content. We can exploit regularities in media and metadata created by communities of users that share common spatial, temporal, and social contexts (as well as other metadata resources such as calendars, contacts, and other contextual knowledge resources) to infer media content from capture and use contexts.

We have developed a camera phone image annotation system that offers unique opportunities for capturing and inferring 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 [5, 6, 13, 17]. A fundamental part of this system is an inference engine that leverages the spatial, temporal, and social context of media creation to infer metadata about media content—our studies have shown this new approach to "context-to-content" inferencing performs well in inferring media content from contextual metadata.

In this paper, we discuss related work in content-based image analysis and systems that use contextual metadata to describe image content (Section 2), describe our approach and the system we built ("MMM" for "Mobile Media Metadata") connecting 55 Nokia 3650 camera phones and a metadata server that infers media content semantics from spatial, temporal, and social contextual metadata (Section 3), assess the performance of that system in inferring the location of the subject of photos (Section 4), and discuss future work (Section 5).

## 2. RELATED WORK

In Smeulders et al.'s survey of content-based image retrieval [14], 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 content 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. 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 [14, 8]. 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 color regions. The identification of the regions as a man, ball, and dog, the relationship among them and the location where the ball is being thrown, and the significance of this event to the person taking the picture are all not represented by low-level machine-extracted features.

As described by [14], 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 intensive knowledge-based approach has only really been viable for highly constrained, controlled, and regularized domains such as industrial automation applications. As we shall see below, it is contextual metadata aggregated across many users and contexts which will enable computational systems to bridge the sensory gap in a wider variety of situations, by, for example, being able to differentiate similar appearing images of the Campanile Tower in Berkeley, California and the Campanile Tower in Venice, Italy, and respectively unifying apparently dissimilar images of either tower.

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. However, recent research that attempts to incorporate background knowledge, effectively "semantic context" in the form of commonsense databases, shows promise for reducing the effort and improving the precision of computer-assisted manual annotation [9, 11]. Nevertheless, most previous work in image annotation is still done long after the image has been captured, and having been thus removed from the context of creation, faces the difficulty of imprecise human memory and the unavailability of the capture context for interactive sensing.

Recent work has looked at addressing parts of these problems by automatically incorporating contextual metadata with the image, most noticeably spatial location. Toyama et al.'s research [15] enables users to tag their photos with GPS data and share these geo-coded 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. However, this work only supported annotation and use of spatial context after the time of capture at a desktop PC and also did not attempt to infer additional contextual and content metadata from the location information.

In ubiquitous computing research, researchers have attempted to use location information to infer additional contextual information as well as the 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 and refine location information based on a system of rules and constraints [10]. Unlike this prior work in context-aware and ubiquitous computing, our research aims to utilize and extend context-aware computing methods 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 bridge the sensory and semantic gaps in media content analysis, retrieval, sharing, and reuse.

Recent related work has begun to attempt to infer media content from the context of media capture. We developed our Mobile Media Metadata (MMM) system [5, 6, 13, 17] at the same time as, and independently of, this related work [12, 16] with some

interesting differences. The LOCALE system at Stanford [12] allows devices and users to share location information and labels for photographic images. Like our own work, it uses location to determine what labels other photographs taken in a similar location should have. LOCALE uses free text annotations of location, while in our own research we use a facetted metadata ontology for media description based on our Media Streams work [2], which includes structured semantic descriptors not only for locations, but for people, objects, and activities as well. Vartiainen's research [16], like our own, features a shared semantic ontology for mobile image annotation, but unlike our research does not leverage social, temporal, and spatial contextual metadata to make inferences about media content. To date, no researchers who are leveraging context to infer media content have yet hybridized the emerging contextual metadata approach with content-based image analysis—we are currently working on such research.

In [5, 13] we provided an overview of our MMM prototype and approach; in [17] we described an evaluation of the users' experience with the MMM prototype; and in this paper we relate our contextual metadata approach to bridging the sensory and semantic gaps in multimedia systems, and describe MMM's "context-to-content" inferencing system in more detail as well as an evaluation of its performance.

## 3. SYSTEM DESCRIPTION

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) [8]. We have explored bridging the semantic gap by leveraging system-directed user interaction at the point of media capture in our research on "Active Capture" [3, 4]. With the advent of mobile phones with cameras, we have a new opportunity to capture, infer, and correct/augment descriptions of media content at the time the media is captured. In leveraging the spatio-temporal *context* and social *community* of media creation and use to help bridge the semantic and sensory gaps, we can take advantage of three aspects of image context that are not only automatically available (and algorithmically and semi-automatically refinable) on camera phones, but also that have special salience in most consumer photos: *when* (the date and time of image capture); *where* (the location of the camera when the image was captured), and *who* (who took the image). By choosing temporal, spatial, and social context, we were able to use the existing camera phone and network infrastructure to gather this information for a given user and across groups of users, algorithmically make inferences to guess, refine, and augment related metadata (e.g., the named semantic location where the camera was located, the named semantic location of the subject of the photo, the person depicted in the photo, etc.), and incorporate user interaction to confirm, correct, and add more metadata when needed at the point of image capture.

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 application concepts for mobile media and metadata creation, sharing and reuse (www.sims.berkeley.edu/academics/courses/is202/f03/phone_project/index.html).

The students were asked to take photos and annotate them 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 a database 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' photos and metadata to a single database to facilitate sharing and correlation of information. For example, if when the majority of users has been standing in or near a location in a given CellID and the majority of them took photos of the Campanile at UC Berkeley, there is a strong chance that if another user is standing in the same spot or somewhere nearby, that they are also taking a picture of the Campanile. By exploiting such regularities in spatial, temporal, and social contexts shared by a network of camera phone users, we were able to leverage the annotative effort of a few users to make inferences about the content of photos taken by a larger community of users. This "context-to-content" inferencing promises to solve the problems of the sensory and semantic gaps in multimedia information systems. For example, 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 spatio-temporal-social context of image capture, we are able to infer that different images taken in the vicinity of the Campanile are very likely of the Campanile at UC Berkeley and know that they are not of, for example, the Washington Monument.

It is important to note that in MMM we are not currently using image processing to determine image content. Rather, we are inferring, and then enabling user verification of, the most probable content of the image by analyzing statistical patterns in prior annotations of images taken at similar times, places, and by related individuals. Furthermore, it is also important to clarify what the "spatial location" of a photo means to our system. For example, if a photo is taken from a vantage point miles away from the Berkeley campus that has a good view of the Campanile, from a user's perspective the photo is "of the Campanile," so the location of the subject of the photo, as opposed to the location of the camera, should be "the Campanile" 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.

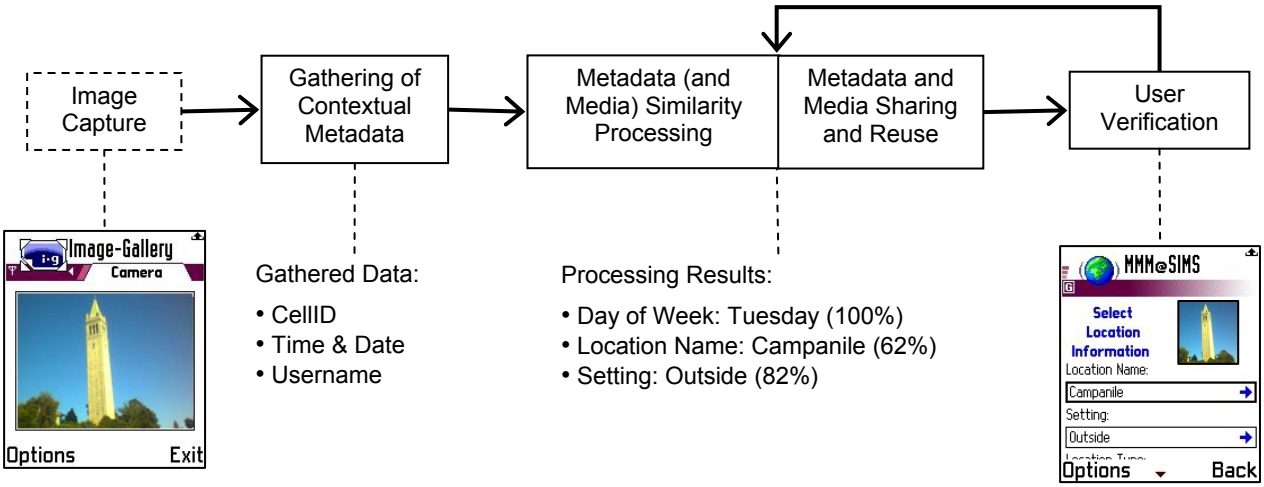


**Figure 1. Mobile Media Metadata image annotation process**

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 (*main subject*, *time*, *date*, network *CellID*, and *username*). Based on this metadata, a server-side metadata similarity algorithm compares the uploaded photo 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 photos and metadata, but contain every user's annotated media to leverage the advantages of shared metadata.

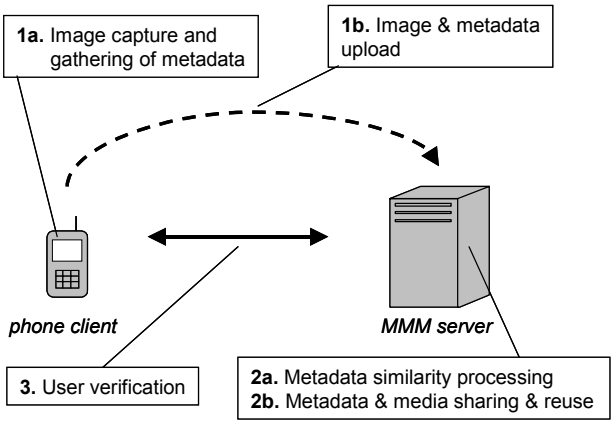


**Figure 2. Mobile Media Metadata system overview**

Using the information from previously captured photos that have similar contextual metadata, the server generates educated guesses (i.e., selection lists with the most probable metadata first) which it presents to the user on the phone's XHTML browser. The user receives the server-generated guesses for verification, and selects or augments the system-supplied metadata. Below we describe the system implementation in more detail by dividing it into the main parts of the metadata creation process.

## 3.1 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* developed in cooperation with Futurice (www.futurice.fi) for the Symbian 6.1 operating system on the Nokia 3650 camera phone. The user captures a photo using *Image-Gallery* which then automatically stores the *main subject*, *time, date,* GSM network *CellID*, and *username*. The image and metadata upload process was implemented in *Image-Gallery* and on the server-side using the Nokia Image Upload API 1.1.

## 3.2 Metadata Similarity Processing

The server-side metadata similarity processing was implemented in a Java module that provides a set of algorithms for inferring metadata for an uploaded photo using the metadata of the uploaded photo and the database of previously annotated photos. 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 and person guessing based on spatio-temporal-social patterns in the contextual metadata using *where* (the phone-supplied CellID and user- and system-refined semantic placename), *when* (the phone-supplied time and date of capture), and *who* (the phone-supplied phone username as well as user-supplied information about the depiction of named individuals in photos). Spatial, temporal, and social context intersect in myriad ways as illustrated in Figure 3.

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 system-supplied and user-supplied contextual and content metadata, the system would 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 locations of photo subjects as well as persons who may have been photographed at a given place and time.

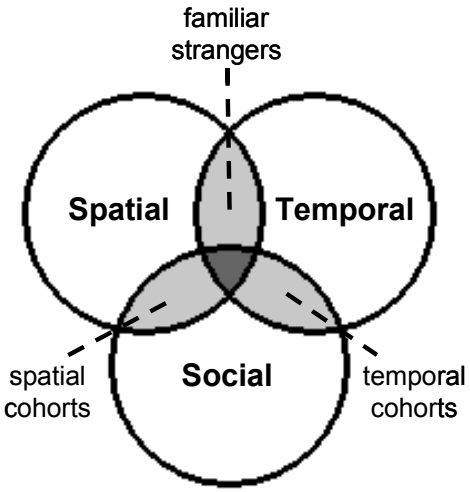


**Figure 3. Spatial, temporal, and social contexts**

MMM uses a weighted combination of spatial, temporal, and social metadata to infer location and person. In its current implementation, it uses a simple linear combination of several features that we determined would be good predictors of image content based on contextual metadata. We chose the simplest implementation for speed and to determine if the features that we had chosen were useful predictors. The location guesser was implemented, used, and evaluated in the MMM prototype and four month trial; the person guesser proved too slow for real-time use and is being redesigned for the next version of MMM.

### *3.2.1 Inferring Spatial Location*

We chose our location-guessing features and weights for them based on our past experience and intuition, and then tuned 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 timeframe (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.

The 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:

1. **Same User** (relative weight 0.6) – assigns high probability scores to locations that have been photographed previously by the same user, and moderate scores to locations in which the current user has been photographed.
2. **Delta(Time)** (relative weight 0.2) – assigns probability scores based on how recently this location has been photographed by the same user.
3. **Same Time of Day** (relative weight 0.1) – assigns high probabilities to locations frequently photographed at the same time of day across all users.
4. **Same Date** (relative weight 0.02) – assigns high probabilities to locations frequently photographed on the same day of the month across all users.
5. **Same Day of Week** (relative weight 0.04) – assigns high probabilities to locations frequently photographed on the same day of the week as the current photo, across all users.
6. **Same Day Class** (relative weight 0.04) – assigns high probabilities to photographs taken on the weekend if the current time is a weekend, or to photos taken on a weekday if the current time is a weekday.

Note that some of the results from the current subalgorithms are mutually reinforcing. For example, a photo taken very soon after a previous photo will score high in all six subalgorithms. In practice, our location-guesser was able to guess the correct location quite effectively. Exempting the occasions in which 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.

## 3.3 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. The photos 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 facetted hierarchical structure. In MMM, the top-level facets were the possible main subjects of the photo: *Person*, *Location*, *Object*, and *Activity*. The objective of the facetted 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, and to utilize orthogonal, recombinable substructures in lower levels of the semantic hierarchy to enable the creation of a large set of structured descriptions from a small set of primitives [2].

While shared metadata is exceptionally useful in inferring media content from context, it is important to recognize the privacy concerns around sharing user profile information with others. Metadata such as time, place, and location all can potentially violate people's privacy. While in our current prototyping environment 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 enabling opt-in/opt-out mechanisms in our systems, aggregating and anonymizing data whenever possible, and are exploring additional means for preserving privacy in future work.

## 3.4 User Verification

The user verification and system responses were implemented in XHTML forms. After uploading the photo and metadata, the client-side *Image-Gallery* program launches the phone's XHTML browser to a URL given by the server during upload. After the server creates the metadata guesses to facilitate the user's

annotation of the image, it creates XHTML pages 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 (as shown in Figure 1).

## 3.5 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 and patterns of times in which a user's prior photos have been 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 [15] can be used to prepopulate the choices of spatial locations. 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 social network services such as Friendster or by harvesting names from a user's already annotated images or contact database. We bootstrapped MMM's ontology by prepopulating the system's ontology 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 system's common ontology, effectively enabling shared bootstrapping. In future work we plan to evaluate the effectiveness of these and other approaches to bootstrapping contextual metadata and shared ontologies.

# 4. SYSTEM EVALUATION

## 4.1 User Studies

In [17] we describe the user studies, surveys, and focus groups we conducted with MMM users. The key findings from these studies were: network speed and unpredictability hamper the use of the phone browser as an interaction interface; for our user population, sharing and browsing photos were more important than search and retrieval; and our users tended to annotate one or two key pieces of information per photo. Below we discuss the performance of the location guesser's "context-to-content" algorithm.

## 4.2 Analysis of Location Guesser Results

We attempted to evaluate the location guesser's performance by measuring it against the probability that its guesses would result from a random sampling of locations within a given CellID. To compute the chance of obtaining results randomly consider a setup with a CellID containing 10 previous photos annotated with 4 unique locations. Let $n_0$=4 be the number of photos annotated with the location of the subject of the photograph under consideration and $n_1$=$n_2$=$n_3$=2 be the number of photos taken of each of the other three locations in the CellID. The probability that the actual location would be selected first is:

$$P(1) = \frac{n_0}{N} = 0.4$$

The probability that it would be randomly ranked second is

$$P(2) = (\frac{3n_1}{N})(\frac{n_0}{N-n_1})(\frac{2n_1}{N-n_1-n_0}) = 0.3$$

(since $n_1$=$n_2$=$n_3$.) Similarly P(3)=0.2 and P(4)=0.1. These computations become increasingly complex as the number of unique locations within a CellID increases, so we compute exact probabilities for cases with fewer than 7 unique locations and use an approximation when there are 7 or unique locations per CellID.

A guess generated by the guesser is assigned a score equal to the probability that a random guess (as above) would be less accurate (rank the actual location lower than the guesser did) minus the probability that a random guess would be more accurate. In the comparison graphs below we also display scores for a hypothetical guesser which ranks locations based on their relative frequency in our entire dataset. Figure 4 shows the overall guesser performance, the ranks assigned by each subalgorithm, results for the relative frequency guesser, and what a perfect score would look like.

In general, the guesser's performance was comparable to that of the relative frequency guesser, though the relative frequency guesser will never be able to guess locations that are statistical outliers, whereas our location guesser is able to guess locations that may be in the general case outliers, but in a particular case correct. While our guesser's performance improves as the system acquires more data, the performance of the relative frequency guesser, and the random performance baseline also improve on this dataset since a few locations occur far more frequently than any others. This coherence and convergence of the dataset are supportive of our approach as a whole: namely that there are statistical regularities in the spatio-temporal-social contexts of both individual and group phototaking that can be leveraged to infer media content. Note also that while each subalgorithm's responses vary wildly, their aggregate (the main guesser's response) does not. The apparent wild fluctuations in the subalgorithms are due, in part, to displaying the ranked results they produce when in fact the main guesser sums the probabilities they return. A subalgorithm could rank a location third, though third and first could be separated by only a few percentage points in terms of the probabilities assigned, thus the graph displayed in Figure 4 tends to enhance variation in the subalgorithm responses.

## 4.3 Suggested Improvements

The simplest way to improve the guesser would be to change the relative weights of each of the subalgorithms. Some preliminary statistical analysis suggests that the Same Date subalgorithm should be weighted more strongly at the expense of the Same User subalgorithm. But this analysis is not exact due to the relatively limited dataset and the difficulty of performing any type of regression when the output of each of the subalgorithms changes with the addition of each datapoint. It might also be possible to assign weightings to the subalgorithms for each user, groups of users, or for each CellID—for example, it's possible that Same Day of Week is a good predictor for some users and not for others.

In future versions of the algorithm we intend to more clearly distinguish and compare the spatio-temporal-social contexts of individual users vs. various groups of users to make better use of

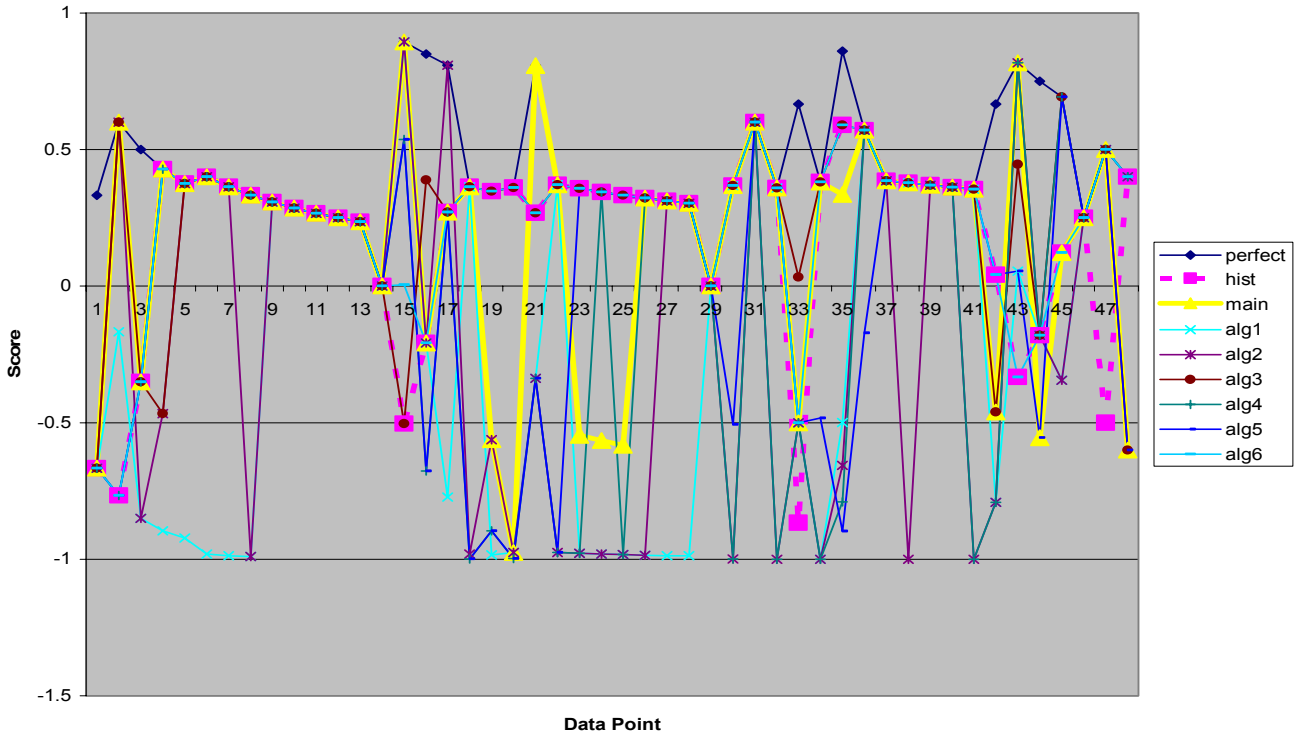


**Figure 4: Location Guesser Algorithm Performance**

the differences and coherences of individual and group phototaking behaviors. The current algorithm mixes individual and group metadata in a way that does not allow us to easily develop, factor, and tune subalgorithms to account for and leverage these differences and similarities.

An additional area for improvement is in more clearly distinguishing between the camera location and the location of the subject of the photograph. There are several examples in our dataset where the guesser performs poorly when a photographer takes two photos very close in time of different subjects. It's likely that the photographer didn't move in space, but merely pointed the camera in another direction, but the Delta(Time) subalgorithm weights the location of the first photo very strongly and suggests that same location for the second photo, discounting the possibility that, though the camera is in the same location, the subject could be quite different. Certain "target-rich" locations will tend toward a high variation in the location of photo subjects within a short timeframe; other locations will have only one or a small number of likely locations for photo subjects. Learning to distinguish these different types of "photo spots" and tuning the subalgorithms appropriately should further improve performance.

It may also be possible to better utilize semantic information about a given location, such as factoring in the type of a location (e.g., educational, commercial, domestic, etc.). Perhaps certain users are more likely to photograph domestic locations on weekends and commercial locations during the week. This would require the output of one subalgorithm (Day Class in this case) to be used as input to another algorithm, which is a significant change from the current system which uses only one "layer" of algorithms where no algorithm's output is input to another.

Finally, in future iterations we plan to explore using additional commonsense knowledge resources, rule-based engines to aid inferencing, and machine learning algorithms to adjust the relative importance of the various location-determining features.

## 5. 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. By leveraging the capabilities of emerging mobile platforms for media and metadata creation, sharing, and reuse and our new paradigm for "context-to-content" inferencing, we have attempted to demonstrate that the prior impeding assumptions of multimedia research could be replaced by a new model for multimedia computing. In this new paradigm we: integrate media capture and media analysis at the point of media creation; leverage spatial, temporal, and social contextual metadata about the capture and use of media across individual users and groups of users to infer media content; and support user-system interaction at the point of capture to enable "human-in-the-loop" approaches to algorithm design.

In our next version of the MMM system, we will be continuing to explore the question of "what did I just take a picture of?" and add to this question the attempt to infer "whom do I want to share this photo with?" In addition to improving our spatio-temporal-social context-to-content inferencing algorithms, we are exploring the integration of contextual metadata and content-based image analysis in order to improve both. In particular, we are working to combine Cognitive Visual Attention (CVA) algorithms with spatial, temporal, and social contextual metadata to better

determine image and metadata similarity as well as validate and refine our contextual information. We also are incorporating additional contextual metadata resources (such as Bluetooth "presence-sensing," personal information management resources such as contact databases, and social network structures and patterns of photo sharing) with face recognition algorithms to better sense, model, and infer co-presence and the likely human subjects of users' photos.

We believe our current promising results and future work in integrating "context-to-content" inferencing and signal analysis will help shape this new and important paradigm for multimedia computing in a way that finally bridges the sensory and semantic gaps in multimedia information systems and enables us to produce multimedia applications that better meet the needs of mobile media users.

## 6. ACKNOWLEDGMENTS

The authors would like to thank British Telecom, AT&T Wireless, Nokia, Futurice, and the Helsinki Institute for Information Technology 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] 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, 3, 3* (Nov. 1996), 179-202.

[2] Davis, M. Media Streams: An Iconic Visual Language for Video Representation. In *Readings in Human-Computer Interaction: Toward the Year 2000*, eds. Baecker, R., Grudin, J., Buxton, W., and Greenberg, S. 2nd ed. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1995, 854-866.

[3] Davis, M. Active Capture: Integrating Human-Computer Interaction and Computer Vision/Audition to Automate Media Capture. In *Proc. of 2003 IEEE International Conference on Multimedia and Expo (ICME2003) Special Session on Moving from Features to Semantics Using Computational Media Aesthetics* (Baltimore, MD, July 6-9, 2003). IEEE Computer Society Press, New York, NY, 2003, Vol. II, 185-188.

[4] Davis, M. Active Capture: Automatic Direction for Automatic Movies. In *Video Proc. of 11th Annual ACM International Conference on Multimedia (MM2003*) (Berkeley, CA, November 2-8, 2003). ACM Press, New York, NY, 2003.

[5] Davis, M. and Sarvas, R. Mobile Media Metadata for Mobile Imaging. In *Proc. of 2004 IEEE International Conference on Multimedia and Expo (ICME2004) Special Session on Mobile Imaging* (Taipei, Taiwan, June 27-30, 2004). IEEE Computer Society Press, New York, NY, 2004.

[6] Davis, M. Mobile Media Metadata: Metadata Creation System for Mobile Images. In *Video Proc. of 12th Annual ACM International Conference on Multimedia (MM2004*) (New York, NY, October 10-16, 2004). ACM Press, New York, NY, Forthcoming 2004.

[7] Dey, A. K. Understanding and Using Context. *Personal and Ubiquitous Computing Journal, 5, 1* (Feb. 2001), 4-7.

[8] Dorai, C. and Venkatesh, S. Computational Media Aesthetics: Finding Meaning Beautiful. *IEEE MultiMedia, 8, 4* (Oct.-Dec. 2001), 10-12.

[9] Haase, K. and Tames, D. Babelvision: Better Image Searching Through Shared Annotation. *ACM Interactions, 11, 2* (Mar.-Apr. 2004), 18-26.

[10] Hull, R., Kumar, B., Lieuwen, D., Patel-Schneider, P. F., Sahuguet, A., Varadarajan, S., and Vyas, A. "Enabling Context-Aware and Privacy-Conscious User Data Sharing. In *Proc. of 2004 IEEE International Conference on Mobile Data Management (MDM'04*) (Berkeley, CA, January 19-22, 2004). IEEE Computer Society Press, New York, NY, 2004, 187-198.

[11] Lieberman, H., Rosenzweig, E., and Singh, P. Aria: An Agent For Annotating And Retrieving Images. *IEEE Computer, 34, 7* (Jul. 2001), 57-62.

[12] Naaman, M., Paepcke, A., and Garcia-Molina, H. From Where to What: Metadata Sharing for Digital Photographs with Geographic Coordinates. In *Proc. of 10th International Conference on Cooperative Information Systems (CoopIS)* (Catania, Sicily, November 3-7, 2003). Springer-Verlag, Heidelberg, Germany, 2003, 196-217.

[13] Sarvas, R., Herrarte, E., Wilhelm, A., and Davis, M. Metadata Creation System for Mobile Images. In *Proc. of Second International Conference on Mobile Systems, Applications, and Services (MobiSYS2004)* (Boston, MA, June 6-9, 2004). ACM Press, New York, NY, 2004, 36-48.

[14] 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, 22, 12* (Dec. 2000), 1349-1380.

[15] Toyama, K., Logan, R., and Roseway, A. Geographic Location Tags on Digital Images. In *Proc. of 11th Annual ACM International Conference on Multimedia (MM2003*) (Berkeley, CA, November 2-8, 2003). ACM Press, New York, NY, 2003, 156-166.

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

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