Eyes on the World

Mor Naaman
Yahoo! Research Berkeley

Contextual metadata can be used to manage large digital photo collections.

It's late April 2011. You've just returned from Paris, where you attended CHI 2011 (conference theme: "Quantum Interfaces: Beyond Fitts' Law"). You turn on the entertainment center (EC) and find your photos from the trip already there—uploaded through a combination of Wi-Fi and cellular connections your camera detected along the way.

You speak into the EC microphone: "Show me that photo of Ben Shneiderman and me in front of la Tour Eiffel," and you also check to see if Ben has seen it. "Show me other peoples' photos taken there at the same time," you tell your EC next; but none of those photos are interesting.

The EC surfaces some other images, though—including several from your honeymoon also taken in front of the Eiffel Tower. "Ah, we were so young," you reminisce as you flip through them. The cyberdog finally arrives with your soy tea. It's disgusting.

LOCATION-AWARE IMAGE CAPTURE

This vision may become a reality well before 2011, with location-aware image-capture devices playing a significant role. Numerous location-aware technologies aim to augment image capture through use of cellular, GPS, and Wi-Fi networks.

Cellular location technology may be the most promising by virtue of being both immediate and ubiquitous. Undeniably, the quality of many camera phones is currently dreadful, but they're rapidly improving. Cell phones are inherently a location-aware technology, so camera-phone photos can easily be associated with location metadata.

The diminishing cost of GPS chips will likely soon enable their inclusion in digital cameras, making them location aware as well. Alternatively, external GPS devices that can be synchronized with a camera are becoming household items; Sony recently released one such device that integrates with its photo browsing system.

Finally, some newer cameras integrate Wi-Fi capabilities and can use technologies like Intel's Place Lab (http://placelab.org) to calculate their location using beacons they detect in "the wild" (this means "large urban areas" for now).

GEOREFERENCED PHOTOS

So now that we know where each photo was taken, what can we do with this metadata? One obvious answer is to plot the images on a map. In 2003, Microsoft Research's World-Wide Media eXchange project (http://wwmx.org) was among the first to do just that. The visually compelling WWMX browser lets users navigate through public photo collections in space and time via a map and timeline interface.

Fast forward to 2006: The popular photo-sharing Web site Flickr (www.flickr.com) recently introduced built-in "geotagging" support that lets users view personal photo collections and public photo pools on a map. Figure 1 shows one such map of images taken in London matching the search term "bridge." Flickr is striving to become the "eyes of the world," an image-based historic archive of cities, neighborhoods, and other regions that can be browsed temporally and spatially.

Extracting useful content

Map-based browsing systems are powerful, but content overload ultimately is unavoidable. In a world where, as Susan Sontag wrote in 1977 (*On Photography*, Farrar, Straus and Giroux), "everything exists to end up in a photograph," map-based presentation tools might have to quickly contend with thousands, and possibly millions, of photos—Flickr already sports close to four million at this time. Luckily, we can mine patterns in these collections to extract meaningful content.

In 1976, social psychologist Stanley Milgram asked his subjects to list places of interest in Paris. Milgram then aggregated the results, effectively creating an "attraction map" of Paris with landmark names appearing in a larger font according to the number of subjects who mentioned each landmark.

We have used the emerging public pools of georeferenced images to extract the same type of information automatically (A. Jaffe et al., "Generating Summaries and Visualization for Large Collections of Geo-Referenced

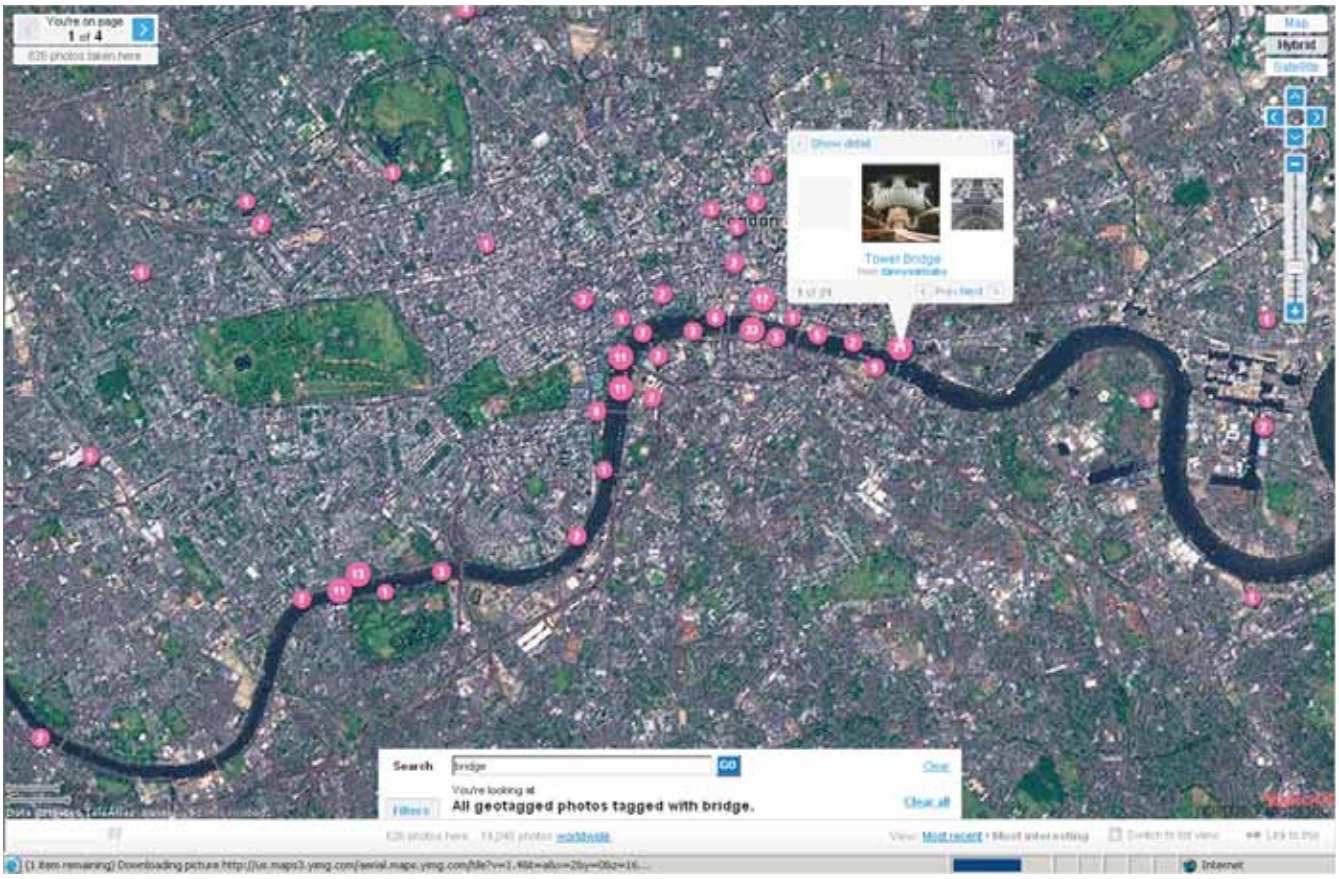

Figure 1. Flickr's geotagging support lets users view personal photo collections and public photo pools on a map—in this case, images taken in London matching the search term "bridge."

Photographs," to appear in *Proc. 8th ACM SIGMM Int'l Workshop Multimedia Information Retrieval*, ACM Press, 2006). The idea is simple: By taking a photo, photographers essentially express their interest in a particular place. Individual pictures taken at a specific location act as "votes" in favor of that location's interest, much like the explicit input of Milgram's subjects.

Further, additional information can be extracted from *tags*, textual labels that users attach to such georeferenced photos on the Flickr Web site. Tags that frequently appear in images from a specific location but are otherwise rare suggest a topic unique to the location.

Tag maps

By analyzing the patterns of location, photographers, and tags in a photo data set, our system can generate *tag maps* that mirror Milgram's manually created attraction map. Figure 2 shows a tag map of central London, derived from Flickr's geotagged photos. The attractions that emerge from the data include Buckingham Palace, London Eye, and Big Ben—all generated automatically with the implicit contributions of Flickr users.

Tag maps are a versatile alternative to previous approaches for visualizing image collections. Moreover, they generally improve as more content is added, alleviating the overload problem often associated with large collections.

Incidentally, the same algorithm can also summarize a photo collection by selecting representative images based on the collection's patterns. The system can then overlay the selected photos on a map in their capture location, effectively illustrating the region's "vibe" via images rather than text.

Support for your own collection

If you can't see how all this data would be useful to your own personal georeferenced photo collection, consider this: If you take a photo near the area marked "Buckingham Palace," chances are good that the photo is of the British monarch's modest residence. At the very least, the palace is a reasonable guess for the photo's content.

Such guesses could be generated from a database of landmarks, but are much more accurate when derived from tags by fellow photographers. Unlike landmark databases, user-supplied tags provide a notion of an object's priority and importance. In addition, user-contributed tags are highly dynamic, changing quickly to reflect new landmarks and attractions.

In fact, tags can even be event-based: "Changing of the guards" could be a relevant label for your Buckingham Palace picture. This phrase doesn't appear in any landmark database, yet Flickr contains dozens of public photos from London with this tag.

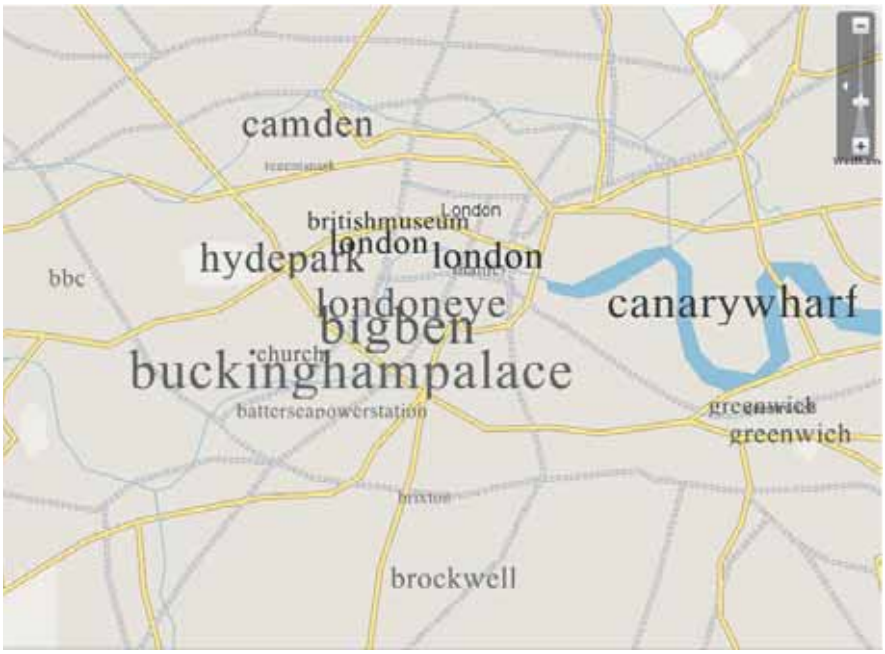

Figure 2. Tag map of London. Attractions are generated automatically from Flickr's geotagged photos.

As the tagging system can now estimate with reasonable accuracy the content of your images, it can help you label the photos (by supplying tag suggestions) or even find photos later without having to annotate them at all.

ZONETAG: A CAMERA-PHONE SOLUTION

Our ZoneTag research prototype (http://zonetag.research.yahoo.com) provides a first look at what these capabilities might afford the user. ZoneTag leverages location information from camera phones (currently Nokia Series 60 phones and Motorola phones like the RAZR), mostly in the form of cell tower IDs that can be translated to a postal/ZIP code or city name.

ZoneTag uploads your camera-phone photos to Flickr, automatically adding place-name tags (such as the city name) so that you can easily search for your pictures by location. Moreover, before you upload your photo, ZoneTag can suggest additional tags that you can attach to the images.

ZoneTag derives its suggestions from tags assigned to your own photos, your contacts' photos, and public photos—all based on their contextual similarity, such as location proximity or proximity in location and time. When you take a picture at home, for example, ZoneTag will suggest the tag "home" and, if appropriate, tags that represent your family members' names, provided you've used those tags in that context before.

For photos taken at work, on a weekday, ZoneTag would suggest work-related tags. You might not even have to type these in yourself: ZoneTag might suggest tags already entered by colleagues for images with a similar context.

And in the case of your British royal palace picture, ZoneTag would be able to suggest both "changing of the guards" and "Buckingham Palace" based on tags used by other people for photos from that location.

ZoneTag has other ways to guess what you're up to when taking pictures. Based on your location, ZoneTag suggests names of nearby venues, restaurants, and even events happening in that area on that day. Whether you're at a Robyn Hitchcock concert at Slim's in San Francisco or having dinner at La Casalinga in Florence, Italy, ZoneTag will be able to suggest these names as tags so you can label your photo as soon as you take it and easily find it later.

If you think ZoneTag has no way of guessing your activity, you can simply point it to an RSS 2.0 feed of personally relevant information. For example, you can point ZoneTag to a list of Upcoming.org events you plan to attend, your personal Web-based calendar data, a list of your favorite picture spots, or names of meeting rooms in your office building. Whenever you take a photo, the relevant feed data will appear in the list of ZoneTag suggestions—the calendar events for that day, the conference room where a meeting is scheduled, and so on—so you can easily tag your image with the appropriate information.

PHOTOCOMPAS: LEVERAGING CONTEXT FOR BROWSING

Location and time metadata are powerful organizational metaphors for images, as Stanford University's PhotoCompas system demonstrates (M. Naaman et al., "Automatic Organization for Digital Photographs with Geographic Coordinates," *Proc. 4th ACM/IEEE-CS Joint Conf. Digital Libraries*, ACM Press, 2004, pp. 53-62).

Consider applications like Picasa (http://picasa.google.com) and Yahoo! Photos (http://photos.yahoo.com) that automatically bind new pictures into location-based folders and event-based albums. No user involvement is required: Import your photos, and you're done.

Because location and time are generally reliable descriptors, your browsing experience can be easily augmented by other images taken nearby at the same time—perhaps by your friends or others at the same event who make their photos public. In fact, a Web-based ZoneTag browser already allows for this type of browsing.

Location and time can also be used to index other types of context. For example, using location and time metadata, PhotoCompas derives weather, daylight status, and more metadata. Using this tool, you can click three times to find the picture you took in Sri Lanka at sunset on that really hot evening. Additional categories derived from location and

time can be added from any existing data source (M. Naaman et al., "Context Data in Geo-Referenced Digital Photo Collections," *Proc. 12th Ann. ACM Int'l Conf. Multimedia*, ACM Press, 2004, pp. 196-203).

COMBINING CONTEXT WITH CONTENT

Although visual features alone are not yet sufficient to reliably identify likely objects, people, places, and landmarks in photos, a combination of context and content can provide a more robust solution. When we know a picture is likely to be either of Buckingham Palace or Big Ben, simple image analysis and indexing techniques might help the system make the final decision.

Similarly, when the context suggests that a person in the photo might be your spouse or child, content algorithms need only consider this constrained list of candidates instead of comparing that person's features to all the people in your collection. Technologies such as Photosynth from Microsoft Live Labs (http://labs.live.com/photosynth) that can analyze a large collection of images of a place or object for similarities and display them in a reconstructed 3D space likewise rely on contextual information.

Indeed, initial research indicates that a picture's context—location, time, event detection, and other factors—is a better predictor of which people are likely to appear in an image than content analysis. Combining context and content features might help alleviate the semantic gap between information extracted from an image's visual features and the human interpretation of that image.

Italo Calvino wrote that a city is described by "relationships between the measurements of its space and the events of its past" (*Invisible Cities*, Harcourt Brace, 1974). Using georeferenced photographs or other media such as video or audio, we can visualize, understand, and utilize these relationships like never before.

Mor Naaman is a research scientist at Yahoo! Research Berkeley. Contact him at http://infolab.stanford.edu/~mor.

Editor: Bill Schilit, schilit@computer.com