Leveraging Geo-Referenced Digital Photographs
Thesis Introduction

Mor Naaman
Stanford University

2005

As the photography world shifted from film cameras into digital cameras, computers now play a significant role in managing people's photographs or, if you will, memories. Photos are stored, shared, searched and viewed — all in digital format.

Managing large personal collections of digital photographs is an increasingly difficult task. As the rate of digital acquisition rises, storage becomes cheaper, and "snapping" new pictures gets easier, we are inching closer to Vannevar Bush's 1945 Memex vision [2] of storing a lifetime's worth of documents and photographs. At the same time, the usefulness of the collected photos is in doubt, given that the methods of access and retrieval are still limited. With digital photos, the opportunity to go "beyond the shoebox" is attractive, yet still not entirely fulfilled.

One of the major hurdles for computer-based photo applications is the *semantic gap.* The semantic gap is defined by Smeulders et al. [8] as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation." Given perfect semantic knowledge about the photos, the task of organizing and retrieving from a photo collection would be made much easier. For example, if a system could automatically derive that a photo shows "Kimya drinking with Dylan at Robyn's birthday, in New York", this semantic knowledge could go a long way in helping users manage and retrieve from their collection. Sadly, current technology sometimes cannot even reliably detect that there are two people in the photo just described.

The existing approaches towards photo collection management can be categorized into four main thrusts. First, there are tools to enable and ease manual annotation (e.g., [7]). These tools let the user rapidly enter semantic information about the photos, to be used later when viewing or searching the collection. However, annotation is still cumbersome and time-consuming for consumers and professionals alike.

Closer to our approach, some systems (such as [5]) attempt to automatically organize photo collections using photo metadata, most notably the timestamps of photos. These systems often supply an interface and easy tools for the users to enhance and improve the organization manually.

The third approach encompasses methods like zoom and pan operations for fast visual scanning of the images (e.g., [1]). These tools attempt to bypass the semantic gap obstacle by posing the personal collection problem as one of visual search. The zooming thus allows the user to find relevant photos without the system having to discern the semantic content ahead of time. The visual tools, though, may not scale to allow the user manage tens of thousands of images without significant semantic information about the photos.

Finally, other systems attempt to directly address the semantic gap using content-based tools that try to extract semantic information from the visual image (refer to [10] for a survey of the area). These tools are not yet, and will not be in the near future, practical for semantic interpretation of personal photo collections. Indeed, low-level visual features can be easily extracted from the images. However, the semantic gap between identifying the low-level features and recognizing important semantic themes in the photographs, is still wide. For example, reliable face recognition is still not available, although recent improvements show better performance with relaxed requirements (e.g., when faces are directly aligned to the camera). Even more farfetched is the ability to identify semantic themes (such as events or activities) by analyzing visual features.

We expand on these areas and more related work in Chapter 2 and in other chapters that are directly relevant to the specific topic.

In the research reported upon in this thesis, we utilize photo metadata such as time and location to help narrow the semantic gap in digital photo collections.

Location is one of the strongest memory cues when people are recalling past events [11]. Location information can therefore be extremely helpful in organizing and presenting personal photo collections. Lately, technology advancements such as Global Positioning System (GPS) and cellular technology made it feasible to add location information to digital photographs, namely the exact coordinates where each photo was taken. While location-aware cameras are not widely available at the time of writing of this thesis, we project that they will become more common in the future. Even today, cameras that can be extended with a plug-in GPS device are available. Other cameras support a GPS API when connected through an external cable. More significantly, cameras embedded in cellular phones are now abundant — cellular is a location-aware technology whose location accuracy will be rapidly improving in the next few years. There are additional ways to produce "geo-referenced photos" using today's technology. For a summary, see Toyama et al. [9]. It is our conviction that future readers of this thesis will find the discussion of location-aware camera technology redundant.

We use time metadata in concert with the location metadata described above. All digital cameras available today embed a timestamp, noting the exact time each photograph was taken, in the photo file's header.[1] The time information is already utilized by commercially available photo browsers (Picasa, iPhotos, Adobe Photoshop Album and others). Novel research systems ([3, 4, 6] and more) also utilize the timestamps, perhaps more aggressively. We discuss those in more detail in Chapter 2.

Given time and location metadata, this research explores various paths to bridging, alleviating or evading the semantic gap in personal collection. Firstly, Chapters 3 and 4 investigate how automatic organization of a photo collection can assist the browse and search tasks. Chapters 5 and 6 look at integrating information from other sources, including user input. In Chapter 7 we expand our settings to allow sharing of information between different users. The discussion explores what additional benefits could be harvested from this type of sharing.

Next, we provide more details on the various parts of this thesis. For each chapter, we note the chapter's contribution to moderating the photo collection's inherent semantic gap.

In Chapter 3 we describe a set of algorithms that execute over a personal collection of photos. Our system, *PhotoCompas*, utilizes the time and location information embedded in digital photographs to automatically organize a personal photo collection. PhotoCompas generates a meaningful grouping of photos, namely browseable location and event hierarchies, from a seemingly "flat" collection of photos. The hierarchies are created using algorithms that

[1] The industry-standard EXIF header supports time as well as location (coordinate) fields.

interleave time and location to produce an organization that mimics the way people think about their photo collections. In addition, the algorithm annotates the generated hierarchy with meaningful geographical names. In Chapter 3 we also test our approach in case studies using three real-world geo-referenced photo collections. We verify that the results are meaningful and useful for the collection owners.

In Chapter 4 we perform a task-based evaluation of PhotoCompas and compare its performance to that of a map-based application. We constructed a browser for PhotoCompas that employs no graphical user interface elements other than the photos themselves. The users interact with the system via textual menus, created based on the automated organization of the respective photo collection into clustered locations and events. The application we compare against, WWMX, features a rich visual interface, which includes a map and a timeline. WWMX is a third party implementation [9]. We conducted an extensive user study, where subjects performed tasks over their own geo-referenced photo collections. We found that even though the participants enjoyed the visual richness of the map browser, they surprisingly performed as well with the text-based PhotoCompas as with the richer visual alternative. This result argues for a hybrid approach, but it also encourages textual user interface designs where maps are not a good choice. For example, maps are of limited feasibility on hand-held devices, which are candidates for replacing the traditional photo wallet.

Chapter 5 introduces a way to help alleviate the semantic gap by adding additional context information about the photo. The idea is that the context in which the photo was taken can sometimes suggest the content of the photo, or at least provide clues for finding photos in a collection. Fortunately, given time and location information about digital photographs we can automatically generate an abundance of related contextual metadata, using off-the-shelf and Web-based data sources. Among these metadata are the local daylight status and weather conditions at the time and place a photo was taken. These context metadata have the potential of serving as memory cues and filters when browsing photo collections, especially as these collections grow into the tens of thousands and span dozens of years. For example, a user may remember that a certain photo was taken during a rainstorm. Even if the rain may not be part of the content of the photo (say, the picture was taken indoors), the context may help the user retrieve the relevant photograph.

We describe the contextual metadata that we automatically assemble for a photograph, given time and location, as well as a browser interface (an extension of the interface used in Chapter 4), which utilizes that metadata. We then present the results of a user study and a survey that together expose which categories of contextual metadata are most useful for recalling and finding photographs. We identify among still unavailable metadata categories those that are most promising to develop next.

Chapter 5 shows that the identity of the people who appear in photos is the most important category in personal photo collections: collection owners often remember photos by the identity of people in them, and often want to retrieve photos using identity-based queries. Chapter 6 tackles exactly this problem. In the chapter, we aim at determining the identity of people in photos of a personal photo collection. Recognizing people, or faces, in images is perhaps the most famous example of a computer's struggle to bridge the gap between the visual and the semantic. Face detection and recognition algorithms have been a major focus of research for the last 15 years, but they still cannot support reliable retrieval from a photo collection, with high recall and precision. Even in the limited circumstances of a personal photo collection, where the number of "interesting" people does not exceed a few dozens, modern face recognition techniques do not perform well enough. A complicating factor in personal collections is that faces are not always well-aligned. In many photos faces are shown in profile, slanted, tilted or even partially or wholly obscured — a fact that makes even detection, let alone recognition, a difficult task.

The system we describe in Chapter 6 suggests identities that are likely to appear in photos in a personal photo collection. Instead of using face recognition techniques, the system leverages automatically available context, like the time and location where the photos were taken, and utilizes the notions and computation of the event and location groupings of photos, as shown in Chapter 3. As the user annotates some of the identities of people in a subset of photos from their collection, patterns of re-occurrence and co-occurrence of different people in different locations and events emerge. Our system uses these patterns to generate label suggestions for identities that were not yet annotated. These suggestions can greatly accelerate the process of manual annotation. Alternatively, the suggestions can serve as a prior candidate set for a face recognition algorithm. Face recognition accuracy may improve when considering fewer candidates, and when assigning confidence partially based on context. We do not incorporate recognition algorithms in this thesis, and leave it for future work.

We obtained ground-truth identity annotation for four different personal photo collections, and used the annotation to test our system. The system proved effective, making very accurate label suggestions, even when the number of suggestions for each photo was limited to five names, and even when only a small subset of the photos was annotated.

In Chapter 6 we introduce user input, specifically as annotation of identities in photographs. In Chapter 7 we leverage another type of user input, namely free-text captions that users may enter for photos in their collection. We also introduce the concept of implicitly sharing information about photographs between users. Sharing will allow us to build a system that can translate from "where" to "what". Here, the semantic gap between the visual representation and the object in the photo — a building, a geographical landmark, or a geographical feature, for example — is alleviated.

More specifically, Chapter 7 describes LOCALE, a system that allows users to implicitly share labels for photographs. For a photograph with no label, LOCALE can use the shared information to assign a label based on labels given to other photographs that were taken in the same area. LOCALE thus allows (i) text search over an unlabeled set of photos, and (ii) automated label suggestions for unlabeled photos. We have implemented a LOCALE prototype that supports both these tasks. The chapter describes the system, as well as an experiment we ran to test the system on the Stanford University campus. The results show that LOCALE performs search tasks with surprising accuracy, even when searching for specific landmarks.

References

[1] Benjamin B. Bederson. Photomesa: a zoomable image browser using quantum treemaps and bubblemaps. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 71–80. ACM Press, 2001.

[2] Vannevar Bush. As we may think. *The Atlantic Monthly*, July 1945.

[3] Matthew Cooper, Jonathan Foote, Andreas Girgensohn, and Lynn Wilcox. Temporal event clustering for digital photo collections. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 364–373. ACM Press, 2003.

[4] Ullas Gargi. Consumer media capture: Time-based analysis and event clustering. Technical Report HPL-2003-165, HP Laboratories, August 2003.

[5] Andreas Girgensohn, John Adcock, Matthew Cooper, Jonathan Foote, and Lynn Wilcox. Simplifying the management of large photo collections. In *INTERACT '03: Ninth IFIP TC13 International Conference on Human-Computer Interaction*, pages 196–203. IOS Press, September 2003.

[6] Adrian Graham, Hector Garcia-Molina, Andreas Paepcke, and Terry Winograd. Time as essence for photo browsing through personal digital libraries. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, 2002.

[7] Ben Shneiderman and Hyunmo Kang. Direct annotation: A drag-and-drop strategy for labeling photos. In *Proceedings of the International Conference on Information Visualization*, May 2000.

[8] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[9] Kentaro Toyama, Ron Logan, and Asta Roseway. Geographic location tags on digital images. In *Proceedings of the 11th International Conference on Multimedia (MM2003)*, pages 156–166. ACM Press, 2003.

[10] Remco C. Veltkamp and Mirela Tanase. Content-based image retrieval systems: A survey. Technical Report TR UU-CS-2000-34 (revised version), Department of Computing Science, Utrecht University, October 2002.

[11] W.A. Wagenaar. My memory: A study of autobiographical memory over six years. *Cognitive psychology*, 18:225–252, 1986.