Eavesdropping on Storytelling

Margaret M. Fleck
Mobile and Media Systems Laboratory
HP Laboratories Palo Alto
HPL-2004-44
March 17, 2004*

photos, indexing, storytelling

This paper presents the design of a photo display system that eavesdrops on photo storytelling, without explicit intervention by the users. The captured narratives are transcribed by a speech recognizer and analyzed to produce keywords for later image retrieval. The photo display sequence and timing information is captured and used to improve the browsing UI. Results are presented from a set of pilot experiments which suggest the feasibility of building such a system. Key technical challenges are identified.

* Internal Accession Date Only
Approved for External Publication

Eavesdropping on Storytelling

Margaret M. Fleck
Hewlett-Packard Laboratories
1501 Page Mill Rd, MS 1138
Palo Alto, CA 94304-1126
fleck@hpl.hp.com

Abstract

This paper presents the design of a photo display system that eavesdrops on photo storytelling, without explicit intervention by the users. The captured narratives are transcribed by a speech recognizer and analyzed to produce keywords for later image retrieval. The photo display sequence and timing information is captured and used to improve the browsing UI. Results are presented from a set of pilot experiments which suggest the feasibility of building such a system. Key technical challenges are identified.

1 Introduction

People are starting to accumulate large collections of digitized personal photos, through a combination of digital photography and digitization of legacy photographs. Currently, a *basic thumbnail browser* presents the user with a chronological ordering of the collection with a small amount of hierarchical structure, based on when the photos were uploaded and/or scanned in. Newer algorithms use time information provided by the camera to create a true time-based hierarchical structure[17]. It is generally accepted that collections would be easier to use, and more valuable to the users, if the photos were also organized and/or annotated in a way that reflected their content.

Existing photo management systems have attempted to provide content-based organization and annotation, but with limited success. Upload batches can easily be given meaningful names. However, it is cumbersome to manually label individual photos and group them into coherent folders. Some systems make it easier for users to add content information, e.g. drag and drop names from a list[38], use speech input rather than typing [2, 27, 41], capture text associated with photos in email or text documents [24], or create narrative sequences of photos [2, 23]. Some programs try to infer annotations or organization from photo properties[5, 12, 13]. Some extrapolate content information from annotated photos to photos with similar GPS location [29] or similar visual features [23].

The central problem is that none of these approaches provide enough content information to make it easy to find photos in a large personal collection [1] or to remind the user of forgotten facts (e.g. the name of someone in an old photograph). Annotating and organizing photos, or even sending email to a friend describing the photos, is an extra task that users rarely do. Photo labels or captions are rare and tend to be terse. GPS location could suggest a set of likely labels for each photo (e.g. nearby landmarks) but it cannot pin down

[1] Most existing digitized personal photo collections are still small and contain mostly recent photos, making it artificially easy to locate items. Future collections will be much larger

what's really in the picture (e.g. a bicyclist) especially in places where the user takes a lot of photographs (e.g. the home). Algorithms for understanding the photos themselves are still unreliable. At best, they would only be able to textualize information already visible in each photo, not provide supplementary information.

However, when people show their photographs to a friend or relative, they will often talk freely and at length about the contents of the pictures (see e.g. [2, 7, 14, 43]). In addition to giving basic facts (e.g. names and places) or discussing what's shown in the photo, they often tell vivid stories about what was going on when the photograph was taken. If the computer could eavesdrop on such *photo storytelling* sessions, we could gather a wealth of data to feed indexing algorithms, as well as capturing evocative audio clips that may be valued by users.

This paper presents a design for a system to eavesdrop on photo storytelling. It surveys prior work, discusses key technical challenges, and describes results from some pilot experiments.

2 Scenario

To visualize what an eavesdropping system might look like, suppose that Mary has just come back from a trip to England and has uploaded her photos onto a home computer. Her friend Nalini brings her family over for dinner. While the men tend the grill, Mary tells Nalini about the trip, bringing up selected photos on the kitchen display screen. The screen notices the photo-related activity and records their conversation.

A couple weeks later, Mary talks to her mother on the phone. Using a remote photo sharing interface, they look through the photos as they talk about the trip. Mary's 4-year-old son also talks for a bit, giving his grandmother his own perspective on what was important on the trip (e.g. his first carousel ride). Again, the interface records the conversation, synchronized with the pictures.

Mary's mother realizes that Mary's grandmother would also enjoy hearing the story, but she knows that Mary rarely calls her grandmother. So she extracts an appropriate audio clip and emails it to Mary's grandmother. As the audio plays on her personal computer, photos are automatically displayed at the same positions in the narrative as when Mary originally told it.

As conversations are recorded, the eavesdropping system transcribes them using a speech understanding system. Some parts of the conversations are too hard for it to understand. From other parts, however, it extracts a somewhat buggy transcript containing many useful names, places, and descriptions of events. It combines this information with any information available from timestamps, GPS location, and Mary's textual annotations.

At the time it is captured, the recorded information is somewhat redundant, because the users still clearly remember what pictures were taken and the stories behind them. It gains value after a lapse of time. Four years later, Mary no longer remembers the name of the town with the interesting cathedral that her friend might want to visit on her upcoming trip. Because she mentioned this detail while talking to Nalini, the system can remind her.

Moreover, Mary may have forgotten exactly where she stored a photo she remembers taking at Stonehenge. But if she types "Stonehenge" into the user interface, the system returns a handful of photos that were displayed around the time she mentioned Stonehenge in her stories. The photo she wants may be one of these photos. Or it may be a slightly earlier or later photo in the same story. Or it may be a photo nearby in the original chronological sequence.

Mary and her son, now 8, also enjoy listening to clips of stories he told as a small kid. They had both forgotten how frightening a carousel seemed to him back then. It's fun for Christopher to compare how he spoke to what his younger brother sounds like right now. And it makes Mary wonder if it might not be fun

to have a third child.

This system operates almost entirely in the background. The system never asks the user to correct its transcripts. The audio clips are available, but the user is not encouraged to listen to his own voice and get nervous about how often he says "you know." The main visible enhancement from the user's point of view is that he can now retrieve photos using content-related keywords.

3 Types of content-producing behavior

Based on previous studies of photo storytelling and audio capture, we can distinguish three basic activities in which users describe the content of photos:

- Photo annotation for personal use,
- Authoring for asynchronous sharing: composing narrations or other types of supplementary text or audio to be viewed later by others, and
- Storytelling: informally sharing photos with friends who are present (either physically or on the phone).

Previous studies (e.g. [2, 35, 37]) suggest that people only rarely create photo annotations (audio or typed) for personal use. This may be partly the fault of the user interfaces (e.g. typing may be difficult). However, the main problem seems to be that the usefulness of annotations only becomes apparent long after the natural time for creating the annotations. By the time people realize that annotations would have been useful, it's too late to easily capture vivid memories and a daunting pile of unannotated photos has accumulated.

Authoring photo-related materials for asynchronous sharing is relatively well supported by existing text-based document authoring tools; email and web pages are widely used for distribution. Users apparently also take pleasure in using audio (in place of typed text) for narration, commentary, or other sorts of photo enhancement[2, 14]. However, my search on materials posted on the internet suggests that use of audio is still very limited It may be that users feel self-conscious about hearing their own recorded voices and speaking when no one is around to listen. Editing may be difficult because speech is a continuous signal which does not easily support making small corrections. It may simply be that existing tools (e.g. HTML) have inadequate support for audio.

Whether commentary is textual or spoken, authoring polished photo stories is inherently a time-consuming activity. Materials must be composed carefully, because the writer is communicating to a person who cannot interact with them. The final product must look attractive and display well on a different computer setup. This makes authoring primarily suitable for sharing particularly interesting photos or documenting special occasions. Moreover, many authored products (e.g. traditional photo albums [43]) are not actually intended to be free-standing but, rather, depend on a verbal narrative supplied live by the author.

Photo storytelling is the more spontaneous informal activity of holding a conversation about photos, displaying photos in the course of a conversation, or using a verbal narrative to present a set of photos. These conversations are usually spoken, but can also be held via typed chat interfaces[37]. The photos may be dynamically extracted from a shoebox or on-line folder, or they may have been pre-arranged into an album.[43] Photo storytelling forms part of the ordinary fabric of social interactions, used to strengthen social bonds as well as convey information [7, 37]. Because people find it easy to talk when prompted by photos, they are used as a non-threatening way to collect data on child language development[4] and to "break the ice" in anthropology [9] and therapy[10].

Photo stories are more difficult for computers to analyze than annotations or authored materials. When people are addressing posterity or a computer, they tend to speak slowly and carefully. When the audience

is a human, they use normal conversational English: faster and with more false starts. The conversation may occasionally drift away from the photo being viewed. Because the audience can see the same photos as the speaker, he need not describe what is readily visible. Because photos are often shown to close friends and family, often people who participated in the events depicted, the storyteller may not explicitly identify familiar objects and people [7, 10, 14].

The freeform nature of photo storytelling sessions makes them a challenge to capture and analyze. However, they seem to be the type of user activity most likely to efficiently yield large amounts of information about the content of photos.

4 System overview

A complete eavesdropping system would consist of one or more photo-sharing user interfaces which also capture conversations, a media storage system, and a conversation analyzer which analyzes the captured data to extract useful content. This paper will concentrate on the conversation analyzer and the capture-related features of the front-end user interfaces. I will not discuss the media storage system, and will simply assume that it is reliable, allows easy sharing among different household devices, and assigns a unique or probably unique (e.g. MD5 hash) ID to each photo.

Photo storytelling could occur at any reasonably-friendly photo display device. Photos could be displayed on large shared screens (e.g. a computer, TV, or wall display) using some version of a thumbnail browser. They could be shown on e-books or using digitally enhanced paper-based books and prints[1, 15]. Mobile users might share them on cell phones or tiny handheld computers[3]. Remote sharing interfaces (e.g. [31]) might be used to stay in touch with family and friends. In an ideal future world, all such interfaces might be instrumented to capture photo-related conversations.

Conversations captured by the display interfaces (audio plus associated photos) are then sent to the conversation analyzer. It must transcribe the conversations, making use of any other available information about the photos (e.g. GPS location), and then collate information from different conversations involving the same photos. Given a search keyword, the system can then retrieve matching audio tracks and photographs. It can supply audio tracks associated with a specific photograph, or vice versa. It can also supply information about the popularity of specific photos.

We expect the incoming data to be of variable quality. Initially, only some photo-sharing interfaces will be instrumented for capture. Some conversations will have loud background noise or discuss topics irrelevant to the photos on display. The start and end of a conversation may sometimes be cut off. Therefore, the conversation analyzer must be opportunistic: make the most of whatever data happens to be available. Because people tend to tell the same stories over and over again, garbled information may be repeated more clearly on some later occasion.

A typical user might conduct photo-sharing conversations using a number of different devices, and a single photo might be discussed at different times by different family members and friends. Some of the photo display devices (e.g. enhanced paper books) may be capable of capture but not analysis. And content extraction for a given photo will be more reliable if we combined data about it from all sources. However, sharing captured conversations between devices or between users raises significant issues (e.g. privacy) beyond the scope of this paper. Therefore, I will assume that sharing occurs, but leave vague the details of how it is done.

The rest of the paper will discuss the individual steps of the capture and analysis process in more detail.

5 Requirements for capture

The first requirement for an eavesdropping system is to capture good quality recordings of storytelling sessions. Specifically, we need to:

- Get people to talk freely,
- Ensure that the storytelling gets recorded, and
- Obtain good quality audio.

When using standard photo annotation interfaces, the first goal is at odds with the other two. For acceptable quality audio, users must speak into the microphone or attach a clip-on mic. The users must turn audio capture on and off at appropriate times, something they don't seem to think of doing during storytelling [2] or similar activities [21]. Synchronizing audio with photos is often achieved (e.g. [2, 15]) by cutting the audio recording at the point when a new photo is selected, so users must arrange their speech to avoid awkward cutoffs. All of this requires that users pay attention to the computer interface rather than the person they are speaking to, which disrupts the flow of conversation.

A first improvement would be to allow users to speak naturally to one another anywhere in the vicinity of the display, without having to speak into a microphone. A relatively recent range of products, developed for dictation and teleconferencing, use an array of several microphones to separate the voice of a speaker sitting in front of a PC from background noise. Similar microphones exist or could be developed for small portable devices and whole-room systems[33].

A specific problem in the home environment is that the TV, radio, or stereo is likely to be on, ensuring that there will often be relatively loud speech or speechlike (e.g. music) background sound. To cleanly separate the target speech, array microphones must be tuned to accept only sound from sources in front of and near the photo display. Users should be discouraged from placing other sound devices in this area. In the long term, smart sound sources (e.g. digital music) might be able to tell the eavesdropping system what they are playing, so that it can be subtracted out.

Second, conversation proceeds more smoothly if the storytelling session is recorded as a single continuous audio track. Although some people naturally have a lot of pauses in their normal conversational style, many other people talk continuously, not even pausing at major topic boundaries. When two people are having a conversation, one may start talking immediately after the other stops or their voices may even overlap. Continuous recording is the norm when recording conversations for scientific study (e.g. [4]).

In a pilot experiment, I asked 11 colleagues at HP labs (all native speakers of American English) to bring in a collection of their personal photographs and tell me about them. The photos were displayed using a standard thumbnail browser, a web browser, or physical album and printouts. The audio was recorded at 16kHz with a GN netcom VA-2000 array microphone (about $100). For two people viewing photos on a personal computer's screen, we have found that this microphone delivers dependably good audio over a wide range of angles, even a bit beyond its nominal 3ft operating range. [2] The speakers were encouraged to talk to me, rather than to the microphone.

In the first three sessions, I tried to record a single clip for each photo. The constant halting of conversation seemed awkward and the sessions yielded largely short caption-like clips (about 6 minutes per speaker, covering 5-9 photos). I allowed the other 8 speakers to talk freely, cutting the audio only infrequently.[3] This seemed more comfortable and allowed me to quickly record 14–35 minutes of audio for each speaker.

[2] Though only when the PC's sound capture card functions properly.

[3] And that only because I didn't trust my audio capture interface enough to create a single long recording.

The speakers frequently produced long stories and covered a large number of photos (in some cases, several dozen), talking continuously as they moved seamlessly from photo to photo.

Finally, it should be possible to detect when users are interacting with the photo display and automatically start/stop audio recording. In a second pilot experiment, I modified a standard thumbnail browser (Danpei [28]) to automatically start audio recording when there is activity (e.g. mouse clicks) on its user interface and stop recording when activity has ceased for a significant period (e.g. 30 seconds). The system automatically records which photos are displayed in the browser's main window and at what times. Later, the photo sequence can be replayed in time with the audio track, producing a vivid recreation of the story.

Although this automatic recording interface seems extremely easy to use, more work would undoubtedly be required to ensure that it reliably captures the whole of storytelling conversations, without too much other material. For example, it may help to start audio recording slightly before the first UI activity, a feature found on some audio recorders and successfully used in another project in our group[34]. It may help to detect when users touch the screen, e.g. when pointing at parts of the on-screen photograph. When there is no explicit activity on the UI, we might decide how long to continue recording based on whether humans are still nearby and looking at the screen (e.g. using face, eye, and/or heat detectors).

Automatic recording also invades the user's privacy. Although users may like it in the long term, they will undoubtedly have worries at first. At a minimum, the interface must let users selectively delete individual recordings. Discretely indicating when the system is recording (e.g. a small light) might be useful for debugging and for reassuring users. Audio tracks could be deleted after some period of time, saving only the extracted content, unless the user has marked the track for saving or taken some action (e.g. including in email) that suggests it should be saved. Finally, there may also be legal restrictions on capturing from certain types of interfaces, e.g. telephones.

6 The Captured Stories

From the raw data captured by the recording interface, we can derive two types of structured information. The *story structure* consists of the sequence of photos shown on the display, as well as the time interval during which each photo was shown on the display. [4] The display sequence may often follow the original chronological sequence, but will skip over defective or irrelevant photos and occasionally jump to another group of photos on a related theme (e.g. another trip to the same city). The *transcript* contains the words of the story, synchronized (e.g. via embedded markup tags) to the sequence of photos. As Figure 1 illustrates, we can expect automatically generated transcripts to be buggy but contain useful keywords.

From the story structure, we can deduce which photos the users like to talk about, how long they like to talk about each one, and which photos typically follow one another in a story. This information could be fed back into the display user interface, to generate popularity ratings for photographs, group related photos, or bias how often photos are displayed in sharing interfaces (e.g. [34]). Captured story structures could be displayed to the user when he is browsing his collection of photos or authoring a web page. When he is in the middle of telling a new story, the interface might automatically suggest a palette of photos that often follow the ones he has already displayed. When photos are stored on another computer, such predictions could also be used to prefetch photos likely to be displayed in the near future.

Transcript data would be primarily used to support keyword search in the display user interface, augmenting whatever keywords have been supplied by explicit annotation. It could also be used to create links between photos or groups of photos which share keywords (e.g. place names). Although each individual transcript will contain many speech recognition errors, the system can exploit the natural redundancy of storytelling:

[4] On interfaces that display several photos concurrently, we might also need to record geometrical information such as relative sizes of the displayed photos and how much of each photo was obscured by overlapping pictures.

THIS IS THE BRIDGE ACROSS THE AISLE OF RIVER NEAR IOWA CITY BOMB THERE ARE AT THIS BRIDGE OR MAYBE THE NEXT ONE DOWN I CAN TELL REALLY TELL FROM THE PHOTO THERE WAS WHAT'S CALLED A RULER DAM WHERE THE WATER GOES ON TO ME THE DAMN RATHER THAN OVER THE TOP OR DOWN THE SIDE A VERY DANGEROUS THING TO GET CAUGHT IN AND EVERY COUPLE YEARS SOME MAYBE IT WOULD TAKE BOTH DOWN THE RIVER THINKING THEY COULD COPE WITH IT AND IT'S NOT GOING TO THE ROLE WHERE DAN AND KILLED REALLY PREDICTABLE UNDERSTAND WHY PEOPLE KEPT DOING IT

this is the bridge across the iowa river near iowa city um there at this bridge or maybe the next one down i can't tell really tell from the photo there was what's called a roller dam where the water goes underneath the dam rather than over the top or down the side uh very dangerous thing to get caught in and every couple years some idiot would take a boat down the river thinking they could cope with it and get sucked under the roller dam and killed eh really predictable i don't understand why people kept doing it

Figure 1: A description of a photo transcribed automatically (top) and by hand (bottom). Some important words (e.g. "Iowa City", "bridge") are recognized correctly, but others are wrongly identified (e.g. "ruler dam" for "roller dam").

people tend to tell the same stories over and over again. By contrast, when the recognizer fails to correctly identify a word, it supplies words that seem to vary somewhat randomly. So a word that keeps occurring in stories about the same group of photos is probably important and transcribed correctly.

Association of keywords with photographs must make allowances for the fact that storytelling relevant to a photo may not be confined to the period when the photo is actually displayed on the screen. When stories are recorded as continuous audio tracks, the conversational transition from one photo to another may happen slightly before or after the user brings up a new photo on the display. Moreover, information said during the display of one photo (especially where and when it was taken) may actually apply to a whole sequence of related photos.

Again, the system may be able to exploit redundancy in the data. First, collating transcripts from several stories involving the same photos may help clarify which words are actually related to each individual photo. Second, it may not be critical that keyword search retrieve exactly the desired photo. It may be acceptable to produce another photo from the same chronological or story sequence, if the user interface makes it easy to navigate along these sequences to the correct photo.

7 Transcribing stories: goals and background

Obtaining reasonably accurate transcripts is one of the most difficult problems in building an eavesdropping system. It is not essential to produce perfect or even coherent transcripts. Rather, our goal is to create useful data for indexing. This means accurately capturing as many content words as possible, particularly words describing people, objects, places, and events. This paper will consider only recognition of adults who speak native or lightly-accented American English, because there is still only limited work on recognition for children and speakers with strong accents.

Current speech recognition systems can produce very high accuracy transcriptions, but only if the topic of the conversation is limited (e.g. airplane reservations) or if the recognizer is tuned to the individual speaker (e.g. dictation systems). Photo stories can cover a wide variety of topics. Storytelling conversations are likely to involve multiple members of the household, as well as visiting friends and relatives. Inside the home, people seem unlikely to reliably wear smart badges or otherwise identify themselves to the system. Therefore, this task requires large-vocabulary speaker-independent recognition.

For transcribing broadcast news, some large-vocabulary speaker-independent systems have a word error rate

(WER) as low as 20% [18, 39]. The speech in this domain is somewhat careful and large sets of transcribed speech from this domain exist and can be used for training recognizers. Word error rate around 28% have been reported for transcription of voice mail and photo annotations[27, 44, 20]. Word error rates on conversational speech comparable to what we expect in photo storytelling are 36-50%, apparently depending on the specific type of speech and how closely the test data match the training data in topic and speaking style [6, 39, 46].

Despite the high error rates, these recognizers have been used to build successful systems for indexing conversational radio shows [39] and voicemail [20, 44]. Indexing performance is difficult to measure precisely, because measured performance depends on the experimenter's choice of test queries (e.g. longer words may be recognized more reliably than short ones). However, useful content-based indexing can be done even with word error rates as high as 50%, partly because important keywords tend to be repeated more than once[39].

Unfortunately, word error rate is only an approximate predictor of how well the transcripts can be used for indexing. In particular, efficiency considerations limit the size of a recognizer's vocabulary. Since words outside the recognizer's vocabulary cannot be recognized at all, a system with good average performance may fail catastrophically on a story containing many *out of vocabulary* (OOV) words. OOV words often include particularly useful indexing keywords, such as proper names. Phonetic matching can be used to retrieve some audio clips containing an OOV word[8, 26, 45], but the rate of false matches seems to be high, because phonetic matching must be relaxed enough to tolerate low-level recognizer errors and variable pronunciations for words. Therefore, other recent research [6, 36, 46] has looked for ways to add new topic-specific vocabulary and phrases to general-purpose language models.

8 Transcription: pilot studies and future work

Data from 8 of the 11 speakers in my pilot experiment (69 minutes of speech) has been transcribed both by hand and by the Calista speech recognizer from HP Labs Cambridge, based on CMU's Sphinx III recognizer [39, 32]. Calista's language model (i.e. its model of the frequency of words and 2-3 word phrases) was trained on transcribed broadcast news data (largely from the HUB496 and HUB497 corpora). Using this model, the recognizer's word error rate on my pilot data was 60%. Although it does extract enough useful keywords to run a simple indexing demo, this error rate is too high for a real application.

There is clear evidence, however, that the broadcast news language model is not well-tuned for photo storytelling. For example, its 57,000 word vocabulary includes "sharia" but not "pokemon." Some photo stories captured in my pilot experiments made heavy use of unusual proper names (e.g. family members) or specialized vocabulary (e.g. landmarks in Venice). I believe, therefore, that its performance can be improved by constructing language models more appropriate for this domain. We can distinguish four possible types of adaptation:

- (Static) adaptation to a spontaneous conversational style,
- (Static) adaptation to photo storytelling in general,
- (Dynamic) adaptation to an individual or family, and
- (Dynamic) adaptation to a specific photo or group of photos.

Compared to broadcast news, informal conversational speech contains more dialog, more false starts, and more filler phrases (e.g. "you know"). Several corpora of transcribed conversational speech exist and have been used successfully in previous systems [6]. They include the large Switchboard corpus [16] (available from [22]) and several smaller corpora: Switchboard cellular, Callhome English, Callfriend English, KING,

and the Santa Barbara Corpus of Spoken American English (available from the [25, 42]). In total, these provide about 3.5 million words of data. This is small compared to the 130 million words in the broadcast news training set, but probably sufficient to model frequent speech patterns.

Photo storytelling also seems to make frequent use of phrases specific to describing photos, such as "this is a picture of." Moreover, people tend to visit many of the same places (e.g. Venice), take pictures of similar landmarks, have similar hobbies, and document similar events (e.g. births). Therefore, I believe that a large collection of other people's photo annotations will provide a very good model for what a user is likely to talk about. The web provides a convenient source of such data.

To obtain photo stories in bulk, we need to devise queries for search engines (e.g. Google) that will return largely photo stories. A set of different query strings is needed, because search engines may limit the number of results returned for any fixed query (e.g. Google restricts the user to 1000). Queries such as "photo album" seem yield many pages containing only short (or no) captions, as well as instructions on how to create albums. Queries such as "my trip" return only photos on certain topics (e.g. trips but not home renovation). A better approach (adapted from [6]) seems to be giving the search engine phrases (in quotes) that are topic-neutral but common in photo storytelling. For example, "this is me in front of." Using 41 queries of this form, I collected 10 million words [5] of text containing many longer, chatty annotations on a wide variety of topics. I believe that much larger amounts of data can be collected in this way.

Finally, much of the specialized vocabulary in these stories is predictable, given some background knowledge about the user and/or the photographs. If we had a GPS location for the photograph, we could infer some likely vocabulary from a gazetteer or from the captions of photos taken by other people at nearby locations (e.g. via a web search). The speech system could then be primed to recognize these words whenever they occur in our stories, a much safer approach than blindly copying over keywords as in [29].

This simple procedure is unlikely to be reliable for several reasons. Photo stories tend to include background and related information that goes beyond the individual photo currently on display. Changes in topic may not be synchronized with changes in the displayed photo. Only a tiny minority of photos are currently tagged with GPS information. Annotations by other people will not supply the proper names of family and friends, common in photo storytelling. And the size of language models makes fast changes in vocabulary tricky.

Therefore, it may be more reliable to infer likely vocabulary for an entire collection of photos, effectively modelling the habits of the user or family, rather than a single photo. The system might be supplied with a variety of basic models (both vocabulary and pronunciations) reflecting different regions and/or ethnic groups. It might then exploit textual sources of information created by the users, such as photo annotations, email, web home pages, web browser logs, and electronic purchase receipts. This approach is being used by other research groups to adapt recognizers for tasks such as meeting transcription[6, 36, 46].

Assuming that we can assemble sufficient data for doing these types of adaptation and personalization, there are still several types of technical challenges. Large language models (e.g. 100M compressed) require efficient algorithms. Training data must be cleaned up, e.g. markup stripped, irrelevant border material (e.g. on web pages) removed, abbreviations and numbers expanded into their spoken forms[40]. Pronunciations must be found or inferred for new words, especially proper names [11]. Models must be kept up to date, e.g. new family members, new fashions in kids toys. The initial transcriptions may need to be revised, because some new words may appear in textual materials only after their first appearance in conversations.

[5] After html markup has been stripped.

9 The Language of Photo Stories

Thus far, we have treated storytelling transcripts as unstructured sequences of words, all alike. Better algorithms might be possible if we had a better understanding of the form and content of photo stories. Although there is an extensive literature on the general topic of storytelling and narrative, I have not yet found prior work investigating the special issues that arise in our application.

One distinctive characteristic of photo narratives is that speakers often use deictic phrases that refer to the photo itself. For example, "this is a photo of X," "this is the X," or "here we are at X." In the small sample I have collected so far, it seems as if such phrases are typically used to introduce the photo when it is first brought up on the display. Detecting these phrases might let us determine exactly when the conversation has shifted to the new photo.

In photo stories, people may mention key facts about the photo such as where and when it was taken, the names of people in it, and what key event it records (e.g. a party or birth). However, it is not clear how often such information is supplied. The pilot data suggests that some common background information (e.g. the location for a trip) may not be given for every photo, but only occasionally, especially at the start of a coherent group of photos. How often are familiar people and places identified (cf. [14])? When describing when the events took place, do people typically give the month or the season (e.g. "spring") or the main event (e.g. "Christmas")?

Because the photos in a story are often closely related to one another, conversation about one photo may refer back to the description of the previous photo or foreshadow a future photo. So they may use phrases like "this is a photo of the same creek" or "this is the other side of the house," which can only be fully interpreted by reference to the earlier descriptions. This apparently happens even when speakers are supposed to be dictating freestanding captions for individual photographs[30].

Finally, a more ambitious goal would be to detect major topic transitions in the story, e.g. so that we can avoid propagating background information (e.g. time and location) across the transitions. Previous work on topic segmentation (e.g. [19]) has relied primarily on low-level linguistic information such as vocabulary similarity. In photo storytelling, we have access not only to the narrative but also to the chronological and story structures, prior narratives involving these photos, and sometimes GPS locations and explicit annotations. This may make the topic transitions easier to spot.

10 Conclusions

Eavesdropping on photo storytelling is a promising but under-explored paradigm for collecting data that helps users organize their photo collections. My pilot studies suggest that it is feasible to construct such systems. We hope to do so in the near future. Though this application stretches the limits of current speech recognition systems, I believe that the technology will become sufficiently robust over the next few years.

Acknowledgements

This work would not have been possible without the generous help of Beth Logan and Pedro Moreno, who supplied their Calista recognizer, language models, and much helpful advice.

References

[1] Back, Maribeth, Cohen, J., Gold, R., Harrison, S., Minneman, S. (2001) "Listen Reader: an electronically augmented paper-based book." Proc. CHI 2001.

[2] Balabanović, Marko, Lonny Chu, Gregory Wolff (2000) "Storytelling with Digital Photographs," CHI 2000, pp. 564-571.

[3] Bartlett, Joel F. (2000) "Rock 'n' Scroll Is Here to Stay," IEEE Computer Graphics and Applications 20/3, pp. 40-45.

[4] Berman, Ruth A.and Dan I. Slobin (1994) Relating Events in Narrative, Lawrence Erlbaum.

[5] Barnard, Kobus and David Forsyth (2001) "Learning the Semantics of Words and Pictures," International Conference on Computer Vision 2, pp. 408-415.

[6] Bulyko, Ivan, Mari Ostendorf, Andreas Stolcke (2003) "Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures," Proc. Human Language Technology 2003.

[7] Chalfen, Richard (1987) *Shapshot Versions of Life*, Popular Press, Bowling Green Ohio.

[8] Clements, Mark, Scott Robertson, Michael S. Miller (2002) "Phonetic Searching Applied to On-Line Distance Learning Modules," IEEE Signal Processing Education Workshop, 2002.

[9] Collier, John Jr. and Malcolm Collier (1986) *Visual Anthropology: Photography as a Research Method*, University of New Mexico Press, Albuquerque.

[10] Cronin, Órla (1998) "Psychology and Photographic Theory," in Jon Prosser *Image-based Research*, RoutledgeFalmer, London and Philadelphia

[11] Deshmukh, N., J. Ngan, J. Hamaker, J. Picone (1997) "An advanced system to generate multiple pronunciations of proper names" Proc. ICASSP '97.

[12] Duygulu, Pinar, Kobus Barnard, Nando de Freitas, and David Forsyth (2002) "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," Seventh European Conference on Computer Vision, pp IV:97-112.

[13] Flickner, M., H. Sawhney, et al. (1995) "Query by image and video content: the QBIC system" Computer 28/9, pp. 23–32.

[14] Frohlich, D., A . Kuchinsky, C. Pering, A. Don, S. Ariss (2002) "Requirements for Photoware," Proc. CSCW 2002.

[15] Frohlich, D.M., G. Adams, E. Tallyn (2000) "Augmenting photographs with audio," Personal Technologies 4, pp. 205–208

[16] Godfrey, J.J., E.C. Holliman, E.C., and J. McDaniel (1992) "SWITCHBOARD: telephone speech corpus for research and development" ICASSP-92, vol. 1, pp. 517–520.

[17] Adrian Graham, Hector Garcia-Molina, Andreas Paepcke, Terry Winograd (2002) "Time as essence for photo browsing through personal digital libraries," Proc. 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 326–335.

[18] Hauptmann, A.G. and Olligschlaeger, A.M. (1999) "Using Location Information from Speech Recognition of Television News Broadcasts," Proc. ESCA ETRW Workshop on Accessing Information in Spoken Audio, Cambridge, England.

[19] Marti A. Hearst (1994) "Multi-Paragraph Segmentation of Expository Text," Proc. 32nd Meeting of the Association for Computational Linguistics.

[20] Julia Hirschberg, Michiel Bacchiani et al. (2001) "SCANMail: Browsing and Searching Speech Data by Content," Proc. Eurospeech 2001.

[21] Ionescu, Arna, Maureen Stone, Terry Winograd (2002) "WorkspaceNavigator: Tools for Capture, Recall and Reuse using Spatial Cues in an Interactive Workspace," Stanford Univ. Tech. Report TR2002-04.

[22] Institute for Signal and Information Processing, Mississippi State, http://www.isip.msstate.edu.

[23] Kuchinsky, Allan, Celine Pering, et. al. (1999) "FotoFile: a consumer multimedia organization and retrieval system," Proc. SIGCHI Conf on Human Factors in Computing Systems, pp. 496–503.

[24] Lieberman, Henry, Elizabeth Rosenzweig, and Push Singh (2001) "Aria: An Agent for Annotating and Retrieving Images," IEEE Computer July 2001, pp. 1–6.

[25] The Linguistic Data Consortium, Philadelphia, http://www.ldc.upenn.edu

[26] Logan, Beth, Pedro Moreno, Om Deshmukh (2002) "Word and Sub-word Indexing Approaches for Reducing the Effects of OOV Queries on Spoken Audio," HLT 2002.

[27] Mills, Timothy, David Pye, David Sinclair, Kenneth Woods (2000) "Shoebox: A Digital Photo Management System," ATT Laboratories, Cambridge, TR 2000-10.

[28] Morino, Shinji (2003) "Danpei – a Gtk+ based Image Viewer", http://danpei.sourceforge.net

[29] Naaman, Mor, Andreas Paepcke, Hector Garcia-Molina (2003) "From Where to What: Metadata Sharing for Digital Photographs with Geographic Coordinates," Stanford Database Group technical report 2003-37

[30] Katerina Pastra, Horacio Saggion, Yorick Wilks (2003) "Extracting relational facts for indexing and retrieval of crime-scene photographs," Knowledge-Based Systems 16, pp. 313-320.

[31] Pilu, Maurizio (2000) "HP Labs PicShare: A synchronous remote photo sharing system," http://www-uk-hpl.hp.com/people/mp/research/picshare.

[32] Placeway, P., S. Chen, M. Eskenazi, et al. (1997) "The 1996 Hub-4 Sphinx-3 System," Proc. DARPA Speech Recognition Workshop, 1997.

[33] Daniel V. Rabinkin, Richard J. Renomeron, et al (1996) "A DSP Implementation of Source Location Using Microphone Arrays," Proc. of the SPIE, Vol 2846, pp. 88-99.

[34] Rajani, Rakhi and Alex Vorbau, forthcoming report on the Memorynet Viewer.

[35] Rodden, Kerry and Kenneth Wood (2003) "How do People Manage Their Digital Photographs," CHI 2003, pp. 409–416.

[36] Rudnicky, Alexander (1995) "Language Modeling with Limited Domain Data," Proc. 1995 ARPA Workshop on Spoken Language Technology.

[37] Schiano, Diane J., Coreen P. Chen, Ellen Isaacs (2002) "How Teens Take, View, Share, and Store Photos," CSCW 2002.

[38] Schneiderman, Ben and Hyunmo Kang (2000) "Direct Annotation: A Drag-and-Drop Strategy for Labeling Photos," Proc. Intern. Conf. on Information Visualization 2000 (IEEE), pp. 88–95.

[39] Van Thong, Jean Manuel, Pedro J. Moreno, et al. (2002) "SpeechBot: An Experimental Speech-based Search Engine for Multimedia Content on the Web," IEEE Trans. on Multimedia 4/1.

[40] Sproat, R., A. Black, et al (2001) Normalization of Non-standard Words, Computer Speech and Language 15(3) pp 287-333.

[41] Stent, Amanda and Alexander Loui (2001) "Using Event Segmentation to Improve Indexing of Consumer Photographs," Proc. SIGIR 2001, pp. 59–65.

[42] Talkbank Project, http://www.talkbank.org.

[43] Walker, Andrew L. and Rosalind Kimball Moulton (1989) "Photo Albums: Images of Time and Reflections of Self," *Qualitative Sociology*, 12/2, pp. 155–182.

[44] Whittaker, Steve, Julia Hirschberg, et al. (2002) "SCANMail: a voicemail interface that makes speech browsable, readable and searchable," CHI 2002, pp. 275–282.

[45] Young, S.J., M. G. Brown, et al. (1997) "Acoustic Indexing for Multimedia Retrieval and Browsing," ICASSP 1997, vol. 1, pp. 199-202.

[46] Yu, Hua, Takashi Tomokiyo, Zhirong Wang, Alex Waibel (2000) "New Developments In Automatic Meeting Transcription," Proc. Intern. Conf. on Spoken Language Processing 2000.