Speech-Based Annotation and Retrieval of Digital Photographs

Timothy J. Hazen[1]*, Brennan Sherry*[1] *and Mark Adler*[2]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA
[2]Nokia Research Center, Cambridge, Massachusetts, USA

Abstract

In this paper we describe the development of a speech-based annotation and retrieval system for digital photographs. The system uses a client/server architecture which allows photographs to be captured and annotated on light-weight clients, such as mobile camera phones, and then processed, indexed and stored on networked servers. For speech-based retrieval we have developed a mixed grammar recognition approach which allows the speech recognition system to construct a single finite-state network combining context-free grammars, for recognizing and parsing query carrier phrases and metadata phrases, with an unconstrained statistical n-gram model for recognizing free-form search terms. Experiments demonstrating successful retrieval of photographs using purely speech-based annotation and retrieval are presented.

Index Terms: photo annotation, audio indexing, audio retrieval

1. Introduction

With the incredible growth of digital photography over the last decade, comes many new and interesting problems surrounding how people should organize, store, and retrieve their photos. Current methods for describing, indexing and retrieving visual media such as photographs typically rely on manually generated text-based annotations of photographs. An obvious extension to existing text-based systems is to incorporate speech into the annotation and retrieval processes. Towards this goal, we have developed a prototype speech-based annotation and retrieval system for repositories of digital photographs.

As an input modality, speech has several advantages over text. First, speech is more efficient than text. A vast majority of people can speak faster than they can type or write, which can make the process of annotating photos faster. Second, speech input does not require a keyboard or pen-based tablet. Thus, annotations could be recorded on small devices, such as digital cameras, at the time and in the setting that a photograph is taken. Some existing commercial digital cameras already possess this audio annotation capability. Finally, speech is more efficient than graphical interfaces for conveying complex properties. Thus, when retrieving a photograph, it is much easier to specify a set of complex constraints (e.g., when a photo was taken, who took it, what is in the photograph, etc.) within a spoken utterance than within a series of grahpical interface pull-downs menus, check-boxes, and/or text-based search bars.

2. Previous Work

Previous work in the audio indexing field has largely focused on high quality audio and video such as news broadcasts [1, 2]. These tasks are aided by the availability of large vocabulary continuous speech recognition systems that have suitably large amounts of audio and textual training materials to provide for high-fidelity transcriptions of new audio. The transcription of private audio data, such as voice mail, is also beginning to see increased attention [3]. This type of data is more problematic because the audio is typically of lower quality due to its spontaneous generation and the use of lower quality audio recording equipment. The personal nature of this type of data also makes it more difficult for developers to obtain large quantities of training materials to improve their recognition models.

The use of audio indexing technology for photograph annotation and retrieval has been proposed and developed in only a limited number of studies [4, 5, 6]. In general these studies have used commercial-off-the-shelf speech recognition software to replace text-entry for annotation and/or retrieval. User studies have found the inadequacies of the speech recognition technology to be a limiting factor in user acceptance of speech-based annotation tools [7]. More sophisticated multi-modal annotation techniques are also being developed to help mitigate the problems of speech recognition errors during annotation [8].

3. Speech-Based Annotation and Retrieval

3.1. System Overview

Our photo annotation and retrieval system is currently comprised of two basic subsystems, one for annotating photos and one for retrieving. For this work, we assume a scenario in which a person uses a mobile device, such as a digital camera or camera phone, to both take a photograph and record a personalized spoken annotation. The photographs could then be retrieved at a later time from the database using a spoken query.

In our experiments, a Nokia N80 mobile camera phone was used as the mobile device for annotating photographs. To create an initial database of photographs with annotations, test users were provided with the N80 camera phones and given two options for collecting photographs; they could either download existing photographs from their personal collection onto the device or they could use the device to take new photographs.

Given a set of photos, users record verbal annotations for each photograph in the set using an annotation application that is run on the phone. Annotations are recorded using an 8kHz sampling rate. For maximum flexibility, the user is not restricted in any way when they provide these free-form annotations. As an example, the photograph in Figure 1 could be annotated by the user as:

Julia and Pluto at Disney World.

The collected photos are uploaded, with their spoken annotations, from the mobile device to an online server which processes and stores the photos and annotations in a database. In addition to the spoken annotations, ancillary information, or *metadata*, associated with each photograph is also stored in the database. This metadata can include various pieces of information such as the owner of the photograph, the date and time the

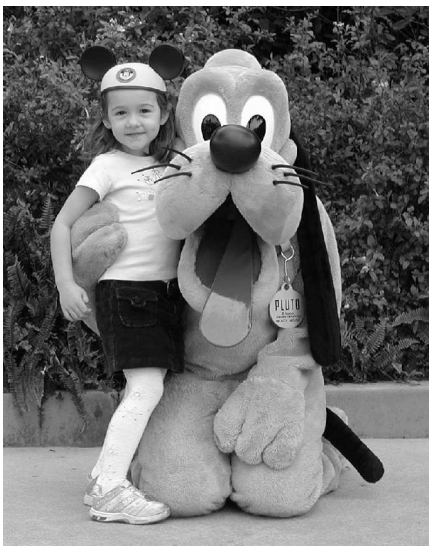

Figure 1: A photograph from the photo annotation database.

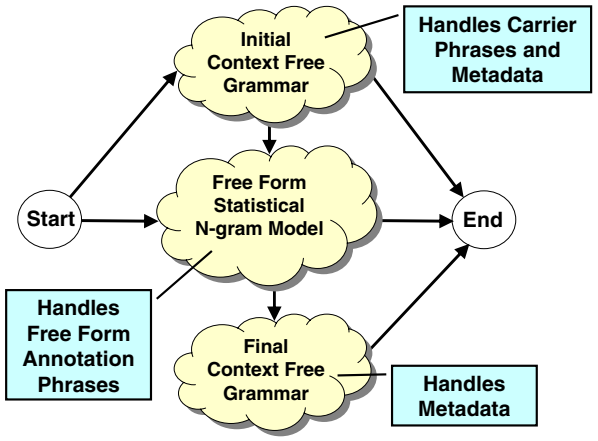

Figure 2: Overview of the flow of subnetworks combined to form the full mixed-grammar finite state network used by the query recognizer.

photograph was taken, and the GPS location of the camera or digital device that took the photograph.

When using the retrieval subsystem, the user can speak a verbal query to specify the attributes of the photographs they wish to retrieve. The system must handle queries that contain information related to either the meta-data and/or the free-form annotations. For example, the user could refer back to the photo shown in Figure 1 with a query such as:

> *Show me John Doe's photo of Julia with Pluto at Disney World from December of 2005.*

This query specifies constraints on both the metadata of the photograph (i.e. whose photograph is it and when was it taken) and the free-form information that describes the contents of the photograph.

3.2. Annotation Processing

Spoken annotations are automatically transcribed by a general purpose large vocabulary speech recognition system. In our experiment, the MIT SUMMIT system was used for recognition [9]. A trigram language model was trained from data obtained from two sources: the Switchboard corpus and a collection of photo captions scraped from the Webshots web page (http://www.webshots.com). The language model training set contained roughly 3.1M total words from 262K Switchboard utterances and 5.3M words from 1.3M Webshots captions. A vocabulary of 37665 words was selected from the most common words in the training corpus. The acoustic model was trained from over 100 hours of speech collected by telephone-based dialogue systems at MIT.

Because of the nature of the task, photo annotations are very likely to contain proper names of people, places or things that are not covered by the recognizers modestly sized vocabulary. In addition to these out-of-vocabulary words, potential mismatches between the training materials and the actual data collected by the system would also be expected to cause speech recognition errors. To compensate for potential misrecognitions, alternate hypotheses can be generated by the recognition system, either through an N-best list or a word lattice. The resulting recognition hypotheses are indexed in a term database for future look-up by the retrieval process.

3.3. Query Processing

To handle photograph retrieval queries, such as the one shown above, we need to be able to recognize and parse queries that may contain several different constituent parts. For examples queries may contain initial carrier phrases (e.g. *"Show me photos of..."*), constraint phrases about meta-data information (e.g.*"...taken in December of 2005"*), or open ended phrases referring to content in the annotations (e.g., *"Julia and Pluto at Disney World"*). To handle such phrases, the system must both recognize the words in the utterances and be able to parse them into their constituent parts. To perform this task we have developed a mixed-grammar recognition approach which simultaneously recognizes and parses the utterance.

In our approach the system integrates both context-free grammars and statistical n-gram models into a single search network. The speech recognition network uses constrained context-free grammars for handling the initial carrier phrases and the phrases for specifying metadata. A large vocabulary statistical n-gram model is used to recognize the free form annotation terms. The context-free grammars and n-gram model are converted into finite-state networks and combined within a single search network which allows the system to move between the grammars when analyzing the spoken utterance. Figure 2 shows the network configuration that the system currently uses. For the query utterance introduced in Section 3.1, the words *"Show me John Doe's photo of"* would be handled by the initial context-free grammar subnetwork, the words *"Julia and Pluto at Disney World"* would be handled by the free-form n-gram model, and the words *"from December of 2005"* would be handled by the final context free grammar.

The structure of the final network represented in Figure 2 provides the user with a large degree of freedom in specifying their query. For example, the user could forego the use of a carrier or metadata phrase, and just speak a phrase containing free-form annotation search terms (e.g. *"Julia and Pluto at Disney World"*), or he or she could forego the use of free-form annotation terms and speak only about metadata (e.g. *"Show me John Doe's photos from December 2005."*). In the worst case, if a user speaks a carrier phrase or metadata expression that is not covered by the context-free grammar, this expression would simply be handled by the n-gram model and treated as

free-form search terms.

The query recognizer uses the annotation recognizer as its initial starting point. The statistical n-gram model used by the query recognizer as well as the acoustic models are identical to those of the annotation recognizer. As with the annotation recognizer, the query recognizer can also produce an N-best list of utterance hypotheses to help compensate for recognition errors and hence improve recall when misrecognitions in the top-choice utterance occur.

The SUMMIT speech recognition system uses a finite-state transducer (FST) representation which allows it to transduce the parse structure and lexical content contained in the recognizer's network into a structured output format. In our case, recognized queries are automatically transduced and into an output XML format by the recognizer. For example, the XML representation which would be generated for the example query discussed above is as follows:

<request>
<owner> john doe </owner>
<terms> julia with pluto at disney world <terms>
<month> 12 </month>
<year> 2005 </year>
</request>

3.4. Photo Retrieval

Photos are retrieved using an or-based word search sorted by a weighted TF-IDF (term frequency-inverse document frequency) score. First, all *stop words* (i.e., non-content words such as articles, prepositions, etc.) are removed from the search term list of the query's N-best list. Then, for each word remaining in the list of search terms, the database is queried and a list of photos containing that term is returned. Each search term yields a TF-IDF score for each photo calculated from information in the database. The TF-IDF score for an annotation term in a particular photo is based on the term's frequency in the entire annotation N-best list, and not just the top-choice answer. The returned TF-IDF score for a term for a particular photo is further weighted by the number of occurrences of the term in the query N-best list. This weighting scheme inherently prefers words which exhibit less confusion in the recognition N-best lists. The weighted TF-IDF scores from each term are then summed to produce the total score for each photo, and all retrieved photos are sorted by their final summed score.

4. Experiments

4.1. Data Collection

To test our photo annotation system we collected a database of 594 verbally annotated photographs from ten different users. Nine of the users provided their own personal photographs. On average these nine users annotated 64 of their own photographs. A tenth user was requested to annotate a collection of 18 generic photographs containing easily identifiable places and objects. As described earlier, annotations were collected on a Nokia N80 mobile phone at a sampling rate of 8kHz.

4.2. Experimental Conditions

To experimentally test the retrieval capabilities of our system, seven of the nine users who provided their own annotated photographs were used as test subjects. Each subject was requested to use our web-based photo retrieval system on a personal computer to retrieve photographs via spoken queries. During the experiment, subjects were shown a photograph and requested to speak a query designed to retrieve the photo from the database. For each spoken query, the system displayed a rank-ordered list of the returned photos as well as the top five interpretations of the spoken query as generated by the recognizer. If the system failed to return the photo within the top-five returned photos, subjects were asked to speak new queries of the system until the system returned the photo within its top-five list. Because the users were given feedback on the system's recognition hypotheses for their query, they could determine if the system correctly or incorrectly recognized their query. This provided some information to the user about whether failed queries were the result of the words or phrases they used or were the result of system misrecognitions. After five successive failures, the subject was asked to move onto a new photo.

Figure 3: Example photographs from the Set A experiment.

In total each subject was asked to retrieve 48 photos. The photos were divided into three experimental sets:

(A) 18 generic photos annotated by a non-test-subject user.

(B) 15 personal photos annotated by the test subject.

(C) 15 personal photos annotated by different users.

Example photographs for the generic Set A are shown in Figure 3. For the Set B photographs, there was at least a one month gap in time between when the subjects annotated their photographs and when the retrieval experiments were conducted. This prevented the subjects from having recent memory of the exact words they used in their annotations.

Before the experiment we anticipated that Set A photos would be the easiest to retrieve and Set C photos would be the most difficult to retrieve. Because Set A photos all contain easily recognizable places and objects, we expected that there would typically be agreement between the annotator and the test subjects on the vocabulary items used to describe the photos. We also expected that the generic nature of the photographs would result in annotation vocabulary items that were likely part of the recognizer's vocabulary. We anticipated that agreement between the annotation and the query for the Set B photos would be high, but that the recognition accuracy on the Set B annotations and queries would be lower because of the personal nature of the annotations.

We anticipated that Set C photos would be the most difficult to retrieve because the test subject may not have knowledge of the people or places in the test photographs. To make this set slightly easier to handle, the subjects were provided with

# of Queries	Set	% of Photos Retrieved	
		As Top Choice	In Top 10
On First Query	A	63.6	66.2
	B	24.7	45.5
	C	23.7	44.7
Within Five Queries	A	88.3	90.9
	B	32.5	70.1
	C	32.5	67.5

Table 1: Photograph retrieval results for the three sets of photographs when examining the subjects' first query only or the first five queries.

the name of the owners/annotators of the photos they were instructed to retrieve (though they were not explicitly told they could use this name to aide in the retrieval of the photograph). Photographs that only showed people with little additional visual context were also filtered out of the Set C test photographs.

4.3. Retrieval Results

Table 1 summarizes the results of our speech-based photograph retrieval experiment. Results are presented in terms of the end goal of retrieving a specific photograph from the repository. We consider a query a "success" if the system returns the desired photograph within the top-ten results returned by the system. In this scenario we are primarily concerned with the system's recall performance. This is in contrast to typical search experiments which place a strong emphasis on returning results with high precision. In this experiment, it is possible that many returned photographs are relevant to a user's query, as users often take multiple pictures at specific events or locations. To avoid passing judgment on the relevance of the photos in the database or the list returned for the specific user queries, we do not attempt to calculate or assess precision/recall numbers for this experiment.

In the table, the results mostly matched our preconceived expectations. The Set A photographs were the easiest to retrieve. These photographs were retrieved by the test subject as the system's top choice photo for their first query attempt 63.6% of the time. When provided with multiple queries to find a photo, the subjects were able to retrieve Set A photos within the top-ten list of the system within their first five query attempts over 90% of the time.

The Set B photographs were more difficult for the subjects to retrieve than we anticipated. Subjects were able to retrieve requested photographs from the database within their first five query attempts 70% of the time. This is only marginally better than the 67.5% rate achieved for the Set C photographs. This indicates that user knowledge of the subject matter of the photographs is not playing as significant of a role as we anticipated. We believe speech recognition errors played the dominant role in the user's difficulties retrieving the photographs in Sets B and C, but further analysis is needed to confirm this belief.

5. Discussion

In this paper we have presented a photograph annotation and retrieval system. All components of the described system have been implemented within a client/server architecture which allows users to annotate and retrieve photographs from a lightweight client (i.e., a Nokia N80 mobile). The processing for the speech-based annotation and retrieval is performed by servers available to the client via a wireless network. A web-based retrieval service has also been developed and was used for the retrieval experiments conducted in this study.

Our preliminary experiments have found that speech-based annotation and retrieval of photographs can be used successfully, particularly when the photos contain common places, objects or people, whose names may reasonably be expected to be found in the captions of the large multi-user photo collection we used for language model training. Retrieval of images that are annotated with obscure proper names not found the recognizer's vocabulary is more problematic. To mitigate this problem we plan to investigate the use of phonetically-based out-of-vocabulary techniques for audio indexing [10]. We also plan to integrate methods for characterizing the photographs based on a variety of properties (i.e., date, audio characteristics of the annotations, visual characteristics of the images, etc.) that would allow the user to refine their searches by specifying new constraints based on the initial results returned by the system.

6. Acknowledgements

This research was supported by Nokia as part of a joint MIT-Nokia collaboration.

7. References

[1] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1338-1353, August 2000.

[2] J.-M. Van Thong, P. Moreno, B. Logan, B. Fidler, K. Maffey, M. Moores. "Speechbot: An experimental speech-based search engine for multimedia content on the web," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 88-96, March 2002

[3] J. Hirschberg, *et al*, "SCANMail: Browsing and Searching Speech Data by Content," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001.

[4] R. Srihari, "Use of multimedia input in automated image annotation and content-based retrieval," *Proceedings of SPIE*, vol. 2420, pp. 249-260, March 1995.

[5] T. Mills, D. Pye, D. Sinclair, and K. Wood, "Shoebox: A digital photo management system." Technical Report 2000.10, AT&T Laboratories, Cambridge, 2000.

[6] J. Chen, T. Tan, P. Mulhem, and M. Kankanhalli, "An improved method for image retrieval using speech annotation," *Proceedings of the 9th International Conference on Multi-Media Modeling*, pp. 15-32, Taiwan, January 2003.

[7] K. Rodden and K. Wood, "How do people manage their digital photographs?" *CHI Letters*, vol. 5, no. 1, pp. 409-416, April 2003.

[8] P. Barthelmess, E. Kaiser, X. Huang, D. McGee, P. Cohen, "Collaborative multimodal photo annotation over digital paper," in *Proc. Int. Conf. on Multimodal Interfaces*, Banff, Canada, November 2006.

[9] J. Glass, "A probabilistic framework for segment-based speech recognition." *Computer, Speech, and Language*, vol. 17, no. 2-3, pp. 137–152, April-July 2003.

[10] T. Hori, *et al*, "Open vocabulary spoken utterance retrieval using confusion networks," in *Proc. ICASSP*, Honolulu, April 2007.