Photo Annotation and Retrieval Through Speech

by

Brennan P. Sherry

[S.B. CS, M.I.T., 2006]

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

June, 2007

Author____
Department of Electrical Engineering and Computer Science
June 19, 2007

Certified by____
James R. Glass
Principle Research Scientist
MIT Computer Science and Artificial Intelligence Laboratory
Co-Thesis Supervisor

Certified by____
Timothy J. Hazen
search Scientist
M] gence Laboratory
nesis Supervisor

Accepted by________
Arthur C. Smith
Professor of Electrical Engineering
Chairman, Department Committee on Graduate Theses

Photo Annotation and Retrieval Through Speech

by

Brennan P. Sherry

Submitted to the
Department of Electrical Engineering and Computer Science
June 19, 2007
In Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

In this thesis I 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 various clients, such as mobile camera phones or the web, 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.

Thesis Co-Supervisor: James R. Glass
Title: Principal Research Scientist
Thesis Co-Supervisor: Thomas J. Hazen
Title: Research Scientist

Table of Contents

Chapter 1 Introduction

The incredible growth of digital photography over the last decade has created many new and interesting problems regarding 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, this thesis outlines 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. Most 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. Also, retrieval can be done on a small device with speech without using a keyboard or pen-based tablet, which are often difficult to use on small devices. Finally, speech is more efficient that 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, where was it taken, what is in the photograph, etc.) within a spoken utterance than within a series of graphical pull menus, check-boxes, or text-based search bars.

1.1 Related Work

Previous work in the audio indexing field has largely focused on high quality audio. The "Rough 'N Ready" system developed by J. Makhoul et al recognized the audio from news broadcasts and stored various pieces of information about the broadcasts in a database for retrieval

[1]. Similarly, J.-M Van Thong et al created a search engine for the web which found and indexed audio posted on the web from news and talk shows and created a retrieval engine called "Speechbot" [2]. "SCANMail", a system proposed by J. Hirschberg et al, indexed users' voice mail and created a user interface for retrieving voice mail from keywords. This work uses lower quality audio recognition from users voice mail recordings [3].

Jiayi Chen et al created a system allowing speech annotation of photographs on digital cameras. This system however limited the users ability to tag. They were specifically told that each picture should be annotated with people, location, event, and date. The results of this system's search capabilities were very poor, with precision falling to below .4 when recall is high, and recall below .3 when precision is high [4].

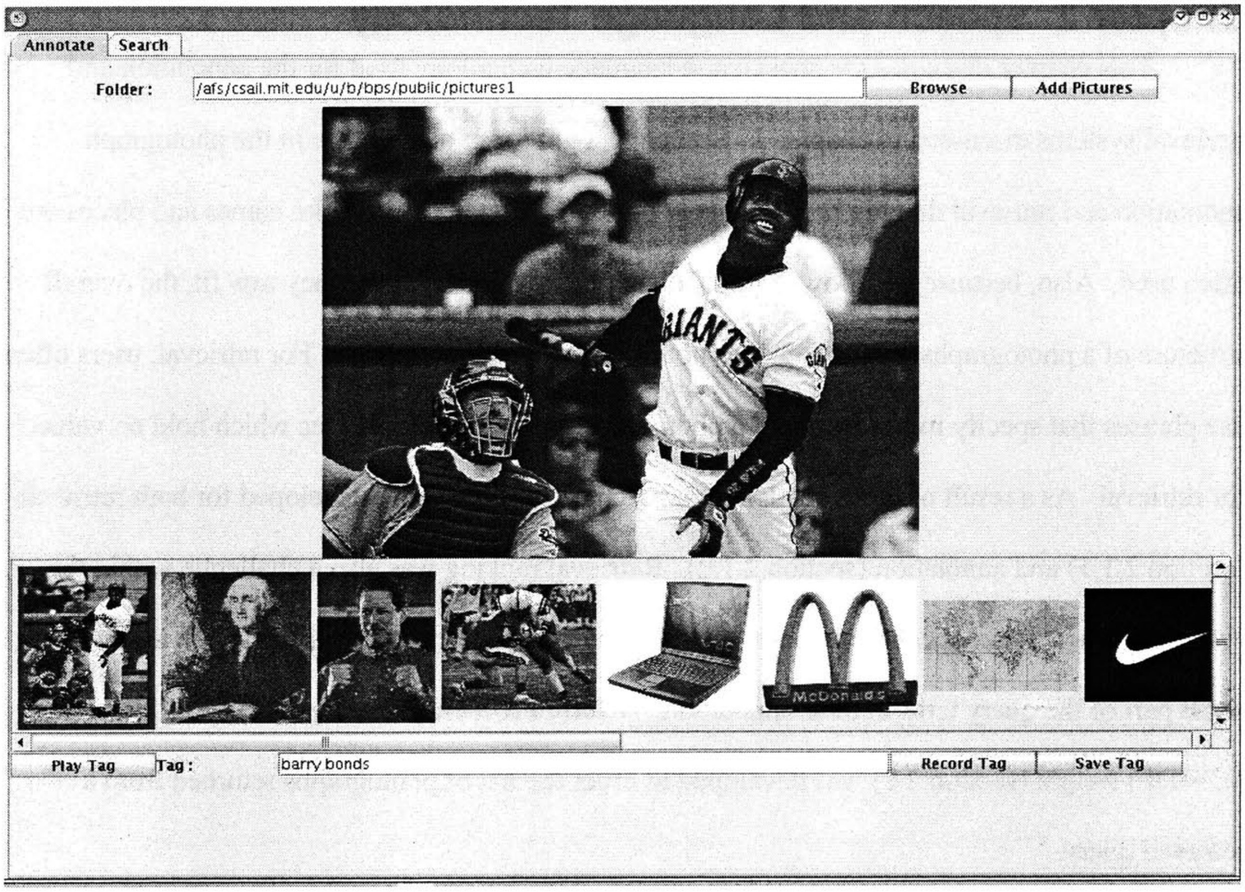
Another system, called PICTION, was developed by Rohini Srihari to identify faces in images. Specifically, the system uses caption processing to specify where faces are in images. The system was however never extended to allow for spoken annotation of the data. It only used premade captions from newspapers to create the recognition process. The system was also not evaluated for its search capabilities [5].

1.2 Background

In spring of 2006, preliminary work was done on this project by the author to show the feasibility of making a system using speech to annotate and retrieve photos. A simple desktop application was developed in Java which had two parts. The first allowed users to annotate photos, shown in Figure 1. The application could display photos, send audio to a recognizer server, and store the recognized n-best list of hypotheses from the recognizer. The second part of the application allowed users to retrieve photos. Users could send audio to the same recognizer, take the resulting n-best list and use its words for retrieval. The system was tested using generic photos which users annotated. Because a special retrieval speech recognizer had not been developed, another set of users annotated the same photos, and those annotations were used as queries with the first set of annoations used as the photographs annotation. Their queries were measured both in terms of their accuracy and precision. Since the recognizer still had a fairly small vocabulary, and because a retrieval recognizer had not been developed, precision and recall were .607 and .439, respectively. While these number were low, it was hoped that improvements to the system would improve retrieval results.

1.3 Scope of Thesis

This thesis outlines the technology behind annotation and retrieval of photographs using speech, as well as systems created for the annotation and retrieval of photographs using speech. Speech recognizers were developed for both annotation and retrieval of photographs, and are discussed in section 2.1. Also, a Term Frequency-Inverse Document Frequency (TF-IDF) scoring method was developed and is explained in section 2.2. Two systems were created for annotating photos. The first, explained in section 3.1, allows users to annotate photos on cellular phones. The second system is a web-based annotation application explained in section 3.2. A web-based application for retrieval was developed and is discussed in section 3.3. The Postgres database which holds photograph and annotation information is explained in section 3.4. Experiments were completed using real users. Both the experiments and the results are discussed in chapter 4. Finally, a summary for the work done for this thesis and the future work for the annotation and retrieval of photographs with speech domain is discussed in chapter 5.



Folder: /afs/csail.mit.edu/u/b/bps/public/pictures1

Tag: barry bonds

Figure 1: A screenshot of the preliminary annotation and retrieval desktop application, created in Spring 2006.

Chapter 2 Speech and Language Technology

This chapter discusses the speech and language technology used for the annotation and retrieval systems discussed in chapter 3. Speech recognition is a challenge in the photograph annotation and retrieval domain because many out-of-vocabulary words like names and places are often used. Also, because we allowed users to annotate photos however they saw fit, the overall structure of a photographs annotation does not follow a specific template. For retrieval, users often use phrases that specify metadata about a photograph, or use carrier phrases which hold no value for retrieval. As a result of these challenges, speech recognizers were developed for both retrieval (section 2.1.1) and annotation (section 2.1.2). Retrieval ranking was also a challenge within the systems. When a user speaks a retrieval query, many photographs are returned which have at least some part of the query term in their annotation. A Term Frequency-Inverse Document Frequency (TF-IDF) weight (section 2.2) was developed to order the list of photographs returned from a retrieval query.

2.1 Speech Recognizers

In this project, the MIT SUMMIT system was used for speech recognition [6]. Two speech recognizers were built for this project. The first was for the annotation system, and the second for the retrieval system. 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 speech recognizer's 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 mis-recognitions, 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 and stored in the system's database for future lookup.

Both an acoustic model and a language model are needed for speech recognition. An acoustic model for the recognizers was trained from over 100 hours of speech collected by telephone-based dialog systems at MIT. A trigram language model was trained from data obtained from two sources: The Switchboard Corpus and a collection of photo caption scraped from the Webshot web Page. (www.webshots.com) The final vocabulary is 37655 words selected from the most common words in the training corpus.

A python script was built to obtain a corpus of captions to train the language model for the speech recognizers. The script would download HTML pages from Webshots and would obtain the caption for a photograph. The filename was then filtered with a dictionary so that any photograph with a random filename given by a digital camera, or any short-hand words, abbreviations, or other non-words were removed. The resulting corpus contained 5.3M words from 1.3M captions.

2.1.1 Retrieval Speech Recognizer

When using the retrieval speech recognizer, 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 metadata and/or the free-form annotations. For example, the user could refer to the photo shown in Figure 2 with a query such as:

"Show me John Doe's photos of Julia with Pluto at Disney World from September 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. To handle photograph retrieval queries the system recognizes and parses queries that may contain several different constituent parts. For example 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.

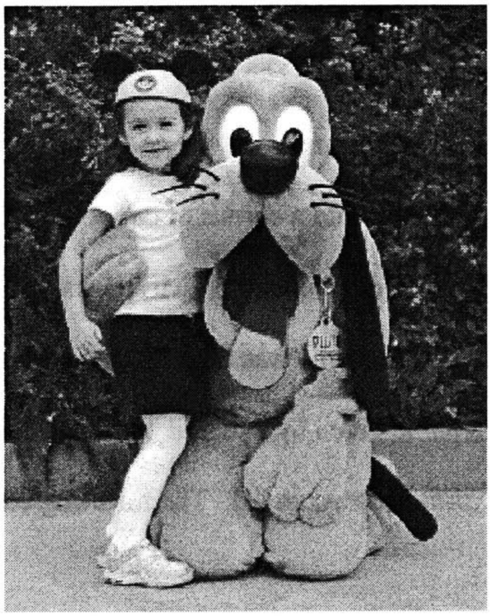


Figure 2: Example photograph used for annotation.

To perform this task the retrieval speech recognizer has a mixed-grammar recognition approach which simultaneously recognizes and parses the utterance. In this 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 3 shows the network configuration that the system currently uses. For the query utterance introduced above, 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 context free grammar.

The structure of the final network represented in Figure 3 provides the user with a large degree of freedom in specifying their query. For example, the user could forgo 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 forgo 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 (section 2.1.2) as it's 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 mis recognitions in the top-choice utterance occur.

The SUMMIT speech recognition system uses a finite state transducer (FST) representation to transduce the parse structure and lexical content contained in the recognizer's network into a structured output format. In retrieval, recognized queries are automatically transduced 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>
```

This XML output is then parsed and used to generate queries into the database.

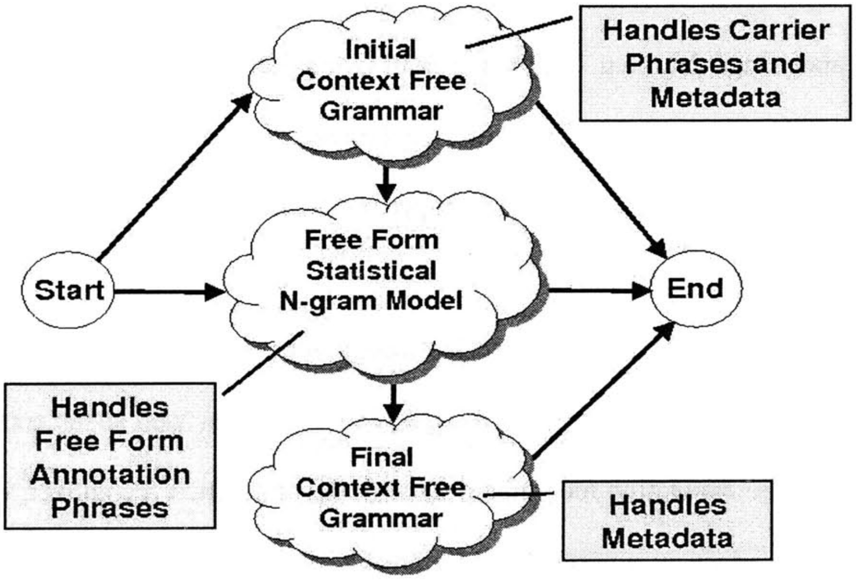

Initial Context Free Grammar
Handles Carrier Phrases and Metadata
Start
Free Form Statistical N-gram Model
End
Handles Free Form Annotation Phrases
Final Context Free Grammar
Handles Metadata

Figure 3: Overview of the flow of subnetworks combined to form the full mixed grammar finite state network used by the query recognizer.

2.1.2 Annotation Speech Recognizer

The annotation recognizer does not need to parse the results of an n-best list into constituent parts. It uses the same language and acoustic models, as well as the same n-gram model as the retrieval speech recognizer. The entire n-best list output is parsed and stored into the database. As a result, the output of the annotation recognizer is much simpler; the n-best hypotheses of what the user spoke, along with the probability of correctness of each entry.

2.2 TF-IDF Scoring

TF-IDF score is a weight often used in information retrieval and data mining to place a value on a word or term used to query information [7]. In this system, a photograph's TF-IDF score is used to sort a set of photographs whose annotations contain words that appear in the query terms of a queries n-best list. To find a photograph's total TF-IDF score, first a TF-IDF score for a photograph for a given word is calculated. To do this, the total number of appearances of a word in the query terms in the n-best list of the photograph's annotation is divided by the total number of photographs which have the word in their n-best list of their annotation. Only words which are not stop words, and appear at least 3 times in the query terms or appear in the top 3 hypotheses of the n-best list of query terms are used. A photograph's total TF-IDF score is the sum of each of its individual word TF-IDF score weighted by the number of appearances of that word in the query terms. The TF-IDF score can be represented by the equation:

$$\text{TF-IDF}_{Pi} = \sum_{\substack{w \in W_{Pi} \bigcap w \in W_{Qj} \\ \bigcap w \in W|(R_w \geq 3||A_{wPi} \geq 3)}} \frac{A_{wPi} \cdot A_{wQj}}{D_w}$$

Where w represents any word which is not a stop word. W_{Pi} and W_{qj} are the set of words in photograph *i*'s annotation and query *j*'s query, respectively. R_w is the maximum rank of the appearance of w in the n-best list of the query. A_{wPi} and A_{wQj} are the number of appearances of word w in photograph *i's* annotation and query *j*'s query, respectively. D_w is the total number of photographs for which word w appears in an annotation.

This technology, both the speech recognizers, as well as the TF-IDF scoring, are used by the annotation and retrieval systems to allow users to quickly, easily, and accurately annotate and retrieve their photos. The overall system architecture is described in the next chapter.

Chapter 3 System Overview

This chapter describes the systems which were created and tested to allow users to annotate and retrieve photographs with speech. Two systems were created to allow users to annotate photos. The first system allows users to annotate photos on Nokia N80 cellular phones (section 3.1). A web-based system was also developed and allows users to annotate photos through a web browser (section 3.2). A web-based retrieval system was developed and allows users to retrieve photographs using speech through a web browser (section 3.3). Common to both systems are a Postgres database (section 3.4) and two speech recognizers (section 2.1). The database stores information on annotations which have been hypothesized by the recognizer. The speech recognizers were developed within SLS at MIT and were created specially for the annotation and retrieval processes.

3.1 Phone-based Annotation System

The first annotation system was developed in python on Nokia N80 cellular phones. Users were given two options for generating photographs; they could either download existing photographs from their personal collection onto the device or they could use the device to take new photographs.

Once photographs had been taken or uploaded, a python script was created to give users a simple UI for annotating photographs. The python used was developed by Nokia for series 60 Symbian phones. In this script, a photograph from the collection is displayed and users can navigate through their photographs one at a time using the directional arrows built into the phone. A menu button allows users to annotate the photographs using the phone's 8 KHz microphone. Once an annotation is recorded, a red banner is displayed around the photograph indicating that an annotation exists associated with that photograph. The script has an option from the menu button to allow users to play back their annotation. If the user is not satisfied with their annotation, they have the ability to rerecord the annotation and the old recording is overwritten. Once annotations have been created for the set of photographs, users have two methods for moving and processing their data so it can be used in the retrieval system.

Figure 4: A nokia N80 cellular phone used for the phone-based annotation system.

3.1.1 Annotations Sent in Batch

In this method, both the photograph and the annotation recording are moved off together and placed into a public folder /(username)/(set number) so that both the photographs and recordings can be easily found later by the retrieval system. A Java program is then run over the photographs to create new, resized photographs for the web annotation system described in section 3.2.2 so that no resizing needs to be done on the fly.

A script is then run over the annotation recordings sending them in batch to the annotation recognizer. The results of the recognition are an n-best hypothesis along with likelihoods associated with each entry. This data is stored in a text file named to associate it with the recording and the photograph.

Once this batch recognition is complete, a python script was developed which creates the entire state of the database into a series of text files, each representing a table in the Postgres database (section 3.4). In this script, the n-best text files created by the batch recognition are parsed, counted, and indexed according to the rules of the tables in the database. Stop words (common words like "the", "is", "a", etc.) are removed for the indexing of words but remain in the stored n-best list so that the full n-best result can be displayed from retrieval if necessary. A Perl script was created to populate the database using these text files.

3.1.2 Annotation Sent in Real Time

Another method for users to move and process their data so it can be used by the retrieval system is to use the python script on the phone to send the data to the system in real time. The python script developed on the phone for annotation has an option in the menu to send the photo and annotation to the server. When this is selected, the script uses XML-remote procedure calls (XML-rpc) and the N80's wifi connectivity to send both the photograph and the annotation recording to an XML-rpc server developed in Java called Venus.

The Venus server accepts the string name of the photo and both the photo file and the annotation recording file and takes all the necessary steps to add it to the system. First it saves both the photograph and the annotation into the correct location (/(username)/(set number)) and then sends the annotation recording to the annotation recognizer. In this case the annotation recognizer is set up as an XML-rpc server. Venus uses XML-rpc calls in Java to connect to the recognizer, send the audio file, and receives back the n-best list of results. This n-best list of results is then parsed by Venus, the stop words removed, and the data is inserted into the database using a Java API developed for interactions with relational databases called Java Database Connectivity (JDBC).

3.2 *Web-Based Photo Annotation System*

A web-based annotation system has also been developed allowing users to annotate photos through a web browser. Users first collect a group of photographs, and they are put in the public file system into a specific folder created for their photographs and annotations (/(user name)/(set number)) The user, knowing their user name and set number, is able to specify their group of photographs in the web-based GUI. The annotation system uses the Galaxy speech system developed within SLS [8]. The various servers used by galaxy for this system are shown in Figure 5.

3.2.1 Web Based GUI

The web-based GUI, shown in Figure 8, allows users to specify a group of photos they wish to annotate by speaker name and set number. Photos are requested using AJAX via a frame request to the GUI manager, a Java server which controls communication between the AJAX on the web page and the rest of the system. The frame request for a group of photos is shown in Figure 6. This request is passed via the Hub to the Annotation Manager. The web page receives an XML response listing the locations of the photographs of the specified group, and displays them.

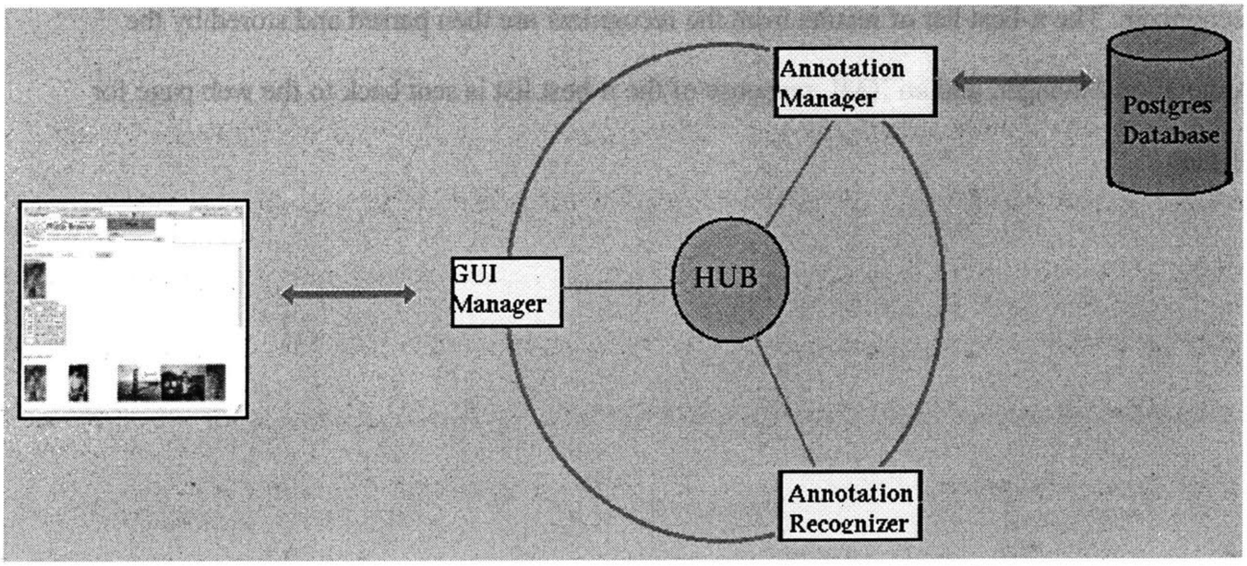


Figure 5: Servers for the Annotation System.

```
{c handle_get_photos :set "setnumber" :speaker "speakername"}
```

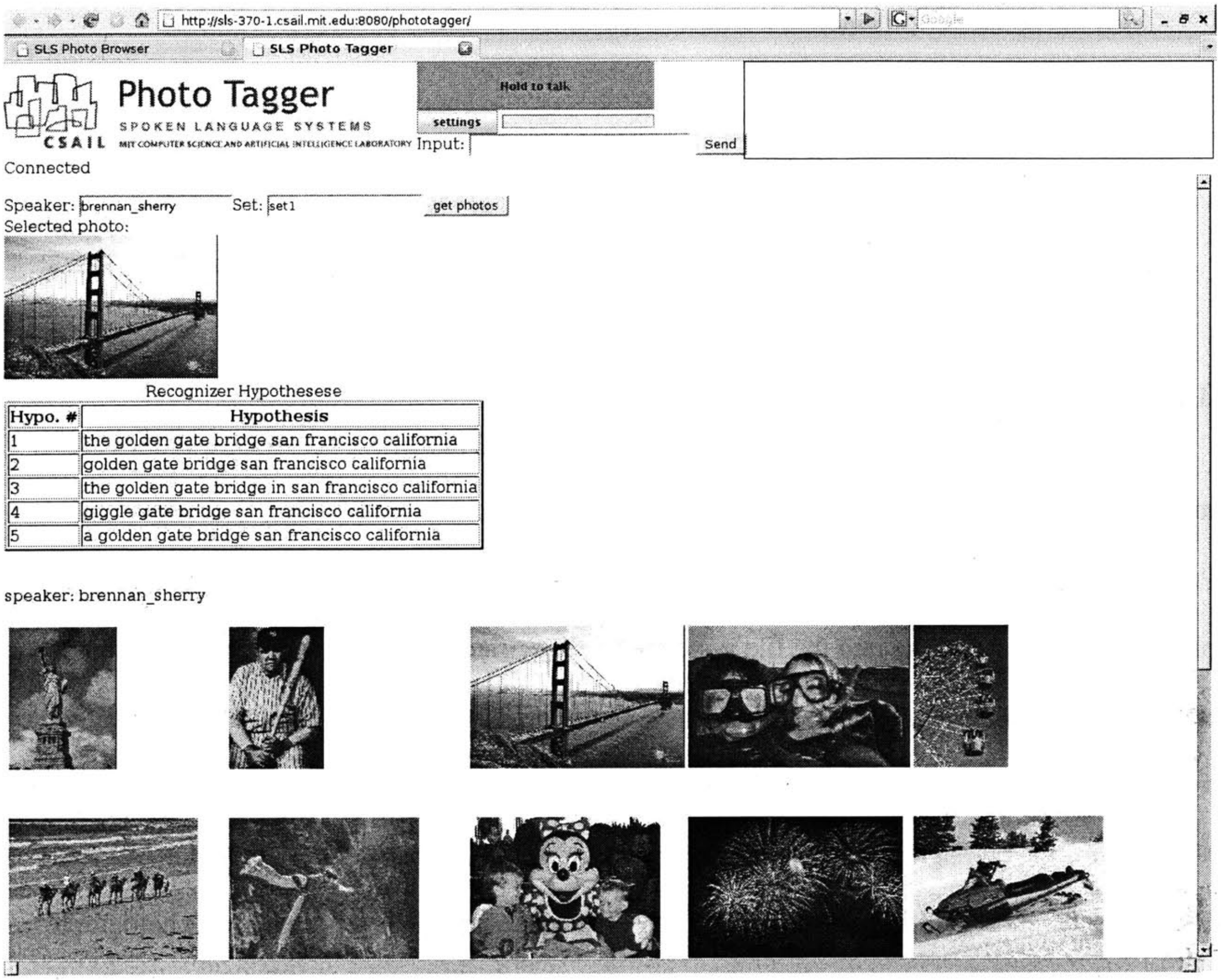
Figure 6: An generic frame for a request of a group of photos.

Photographs can then be selected by the user. When this occurs, if an annotation already exists for the selected photograph, the annotation recording is played and the top 5 hypothesis from the n-best list recognition of the recording is displayed. The n-best list is obtained via a frame request, shown in Figure 7, which is passed to the Annotation Manager via the GUI Manager and the Hub. The web browser receives back an XML response specifying the n-best list of hypotheses.

```
{c handle_get_annotation :set "setname" :filename "filename.jpg"  :speaker
"speakername"}
```

Figure 7: An generic frame for a request of an n-best list for a photo.

A user can record an annotation using a Java servlet which is uploaded when the page is reached. To annotate a photo, a user holds the "hold to talk" button at the top of the page (figure 8), and the Java servlet streams audio from the microphone on the users computer to the annotation recognizer. The n-best list of results from the recognizer are then parsed and stored by the Annotation Manager, and an XML response of the n-best list is sent back to the web page for display.


http://sls-370-1.csail.mit.edu:8080/phototagger/
SLS Photo Browser
SLS Photo Tagger
Photo Tagger
SPOKEN LANGUAGE SYSTEMS
CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY
Hold to talk
settings
Send
get photos

Input:

Connected

Speaker: brennan_sherry Set: set1

Selected photo:

Recognizer Hypothesese

Hypo. #	Hypothesis
1	the golden gate bridge san francisco california
2	golden gate bridge san francisco california
3	the golden gate bridge in san francisco california
4	giggle gate bridge san francisco california
5	a golden gate bridge san francisco california

speaker: brennan_sherry

Figure 8: An example of the web-based Annotation System GUI in use.

3.2.2 Hub

The Hub server manages all communication between servers in the system. Rules are written to specify how any frame passed within the system should be handled. An example of a Hub rule is shown in Figure 9. Specifically, the Hub sets up a port for the Java audio servlet to stream audio to the recognizer. It then accepts an n-best list result from the recognizer and passes it to the Annotation Manager. The Hub also takes any request from the GUI Manager for photo information (location or n-best list) and passes it to the Annotation Manager, and also takes any XML information about a photograph from the Annotation Manager (location or n-best list) and passes it back to the GUI Manager.

```
RULE: :nbest_list --> photomanager.handle_nbest_list
IN: :nbest_list
OUT: :XML_string :selected_input_string :reply_frame
LOG_OUT: :XML_string
```

Figure 9: An example of a Hub rule. Hub rules control the flow of messages and information throughout the system.

3.2.3 Annotation Manager

The Annotation Manager is a Java server in charge of managing the storage of annotations in the system and the retrieval of photograph URL's. The Annotation Manager connects to the Postgres database with JDBC. When the Annotation Manager receives a frame requesting photograph URLs, it looks into the file system at the specified user name and set number location to retrieve the names of all photos. When the Annotation Manager receives a frame requesting the n-best list of a specified photograph, the Annotation Manager creates an XML response holding the n-best list of the photograph by connecting and querying the database. The Annotation Manager also holds the photograph name of the last n-best list requested. This represents the photograph currently selected, and when an n-best list is received from the annotation recognizer via the Hub, the photograph being annotated is known.

When an n-best list is received from the annotation recognizer via the Hub, any n-best list formerly associated with the specified photograph is removed, and all indexes to words in the old n-best list are removed. Once this is completed, the new n-best list is parsed, the stop words are removed, and inserted into the database appropriately. The n-best list is returned to the web browser via the GUI Manager and the Hub in an XML response.

3.3 Web-based Photo Retrieval System

A web-based retrieval allows users to use speech to retrieve photos. It is a Galaxy speech system [8] similar in overall structure to the web-based annotation system. A diagram of servers can be found in Figure 10. Users can specify a query with speech, and audio is streamed to the retrieval recognizer (2.2.1). The results from the recognizer are passed by the Hub (section 3.3.2) to the Photo Manager (section 3.3.3). The Photo Manager queries the database (section 3.4) and returns a list of organized photographs locations for display. This list is passed back to the Hub and is sent to the GUI Manager (section 3.3.1), which relays the result to the AJAX web client front end (also section 3.3.1)

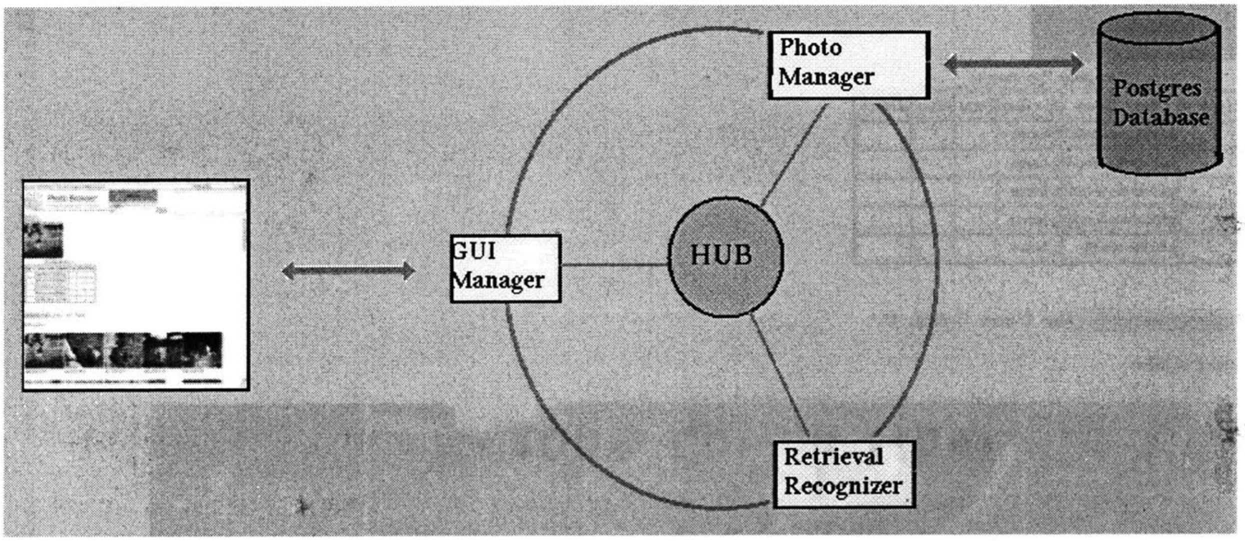
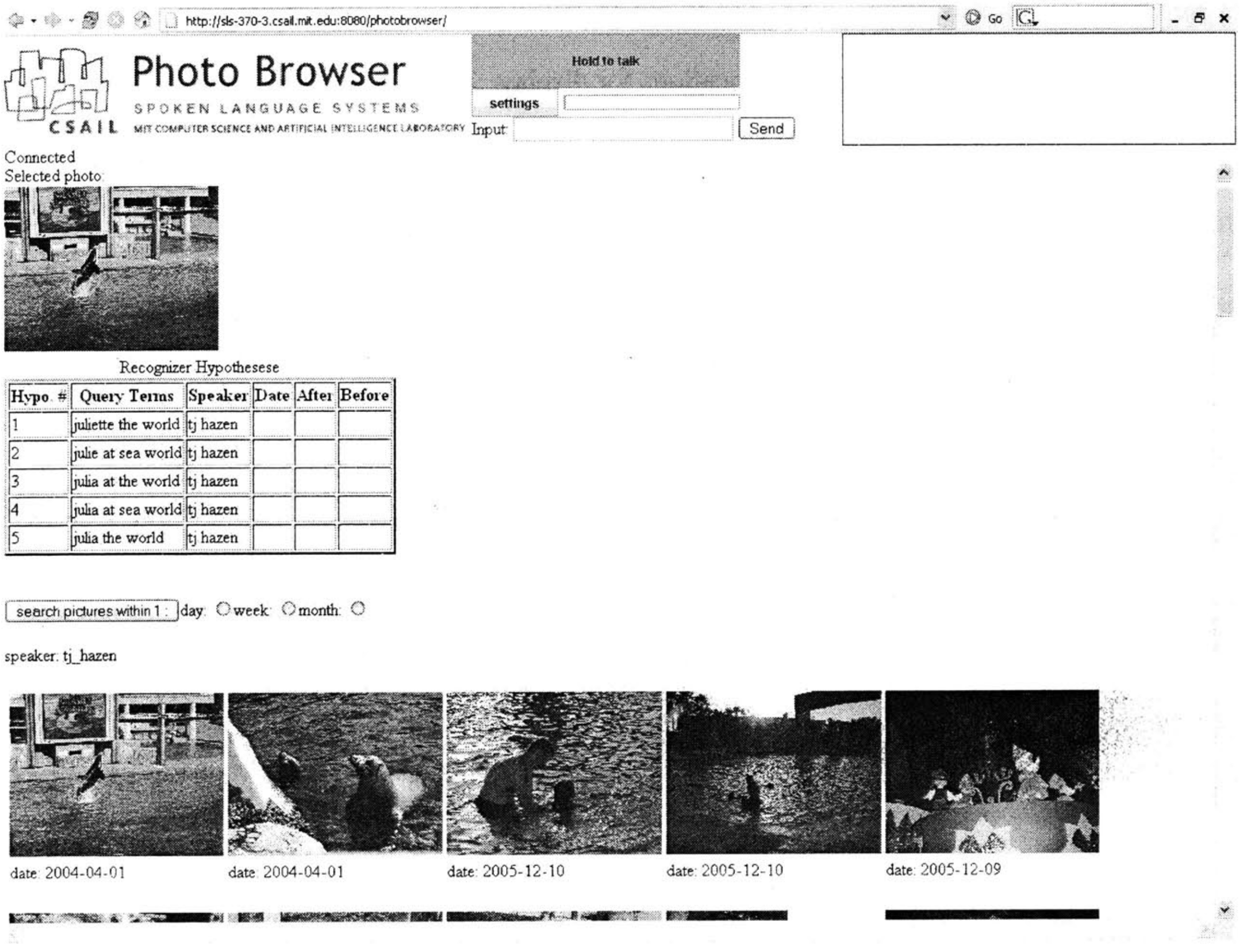


Figure 10: Servers for the Retrieval System.

3.3.1 Web Based GUI

The web-based GUI, shown in Figure 11, allows users to speak to a recognizer and view photographs with annotations and metadata which match their queries. The GUI communicates with the system with AJAX via the GUI Manager. To query the system for photographs, the user holds down the "hold to talk" button at the top of the page (Figure 11). When the user holds this button, a Java servlet within the web page streams audio to the retrieval recognizer (section 2.1.1).

As an, suppose the user spoke "Show me TJ Hazen's photos of Julia at Disney World from December 2005." into the system.


http://sls-370-3.csail.mit.edu:8080/photobrowser/
Go
Photo Browser
SPOKEN LANGUAGE SYSTEMS
CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY
Hold to talk
settings
Input:
Send

Connected
Selected photo:

Recognizer Hypothesese

Hypo. #	Query Terms	Speaker	Date	After	Before
1	juliette the world	tj hazen			
2	julie at sea world	tj hazen			
3	julia at the world	tj hazen			
4	julia at sea world	tj hazen			
5	julia the world	tj hazen			

search pictures within 1 : day: ○ week: ○ month: ○

speaker: tj_hazen

date: 2004-04-01 date: 2004-04-01 date: 2005-12-10 date: 2005-12-10 date: 2005-12-09

Figure 11: An example of the Retrieval System GUI in use.

The web GUI receives an XML response from the GUI manager representing both the n-best list of hypotheses from the recognizer, as well as a list of photograph locations which match the query, as well as dates the photographs were taken. In our example, the GUI would receive all photographs annotated by John Doe, taken during December 2005, whose n-best annotation hypotheses contain at least 1 word from the n-best retrieval hypotheses' query terms. The details of retrieval and ordering of photos are specified in the Photo Manager (section 3.3.3) and TF-IDF scoring (section 2.2).

The web GUI then displays all the photographs returned from the query, the dates they were photographed, as well as the top 5 hypotheses from the n-best list. Once users are displayed photos, they have the ability to select photos. When a picture is selected, it is displayed on the top of the page, and the recording of its annotation is played to the user. This allows users to compare what they had just spoken to what was annotated previously. Also, the user can do date searches on the selected photo by specifying a time period using the buttons at the top of the page and selecting the "Show pictures within 1" button. In the current implementation the user can specify 1 day, 1 week, or 1 month as a time interval.

In a date search, the system returns all photos which were taken within the time interval of the time the selected photograph was taken. As an example, if a selected photo was taken April 4th, 2006, if the system was queried with a date search of one month, the system returns all photographs taken from March 4th, 2006 until May 4th, 2006. The query process occurs in the Photo Manger (section 3.3.3). The web GUI sends a date-query frame to the GUI Manager specifying a photograph and a time interval. The web GUI receives an XML response specifying the locations of the photographs which were taken within the selected interval.

Date queries were implemented to show how metadata can be used not to specify photos with speech, but also in relation to other photos. If GPS coordinates one day are stored in photograph metadata, an easy extension to this system would be interval searches to show photographs taken within x distance of a specified photograph. The query would work very similarly to a date query and could be easily implemented.

3.3.2 Hub

Similar to the web-based annotation system, the Hub server manages all communication between other servers in the retrieval system. Rules are written to specify how frames from the various servers should be handled. The Hub sets up a port for the Java audio servlet to stream audio to the retrieval recognizer. The Hub then takes the n-best list frame from the recognizer and sends the frame to the Photo Manager. The Hub takes an XML result from a spoken query or a date query from the recognizer and passes it back to the web-browser via the GUI Manager.

3.3.3 Photo Manager

The Photo Manager server is the engine which executes queries to the Postgres database to find photographs to display. Written in Java, the Photo Manager accepts both n-best lists and date queries and produces and XML string representing all pertinent information for display.

Date Query: The Photo Manager handles requests for date queries by accepting date-query frames from the Hub. The Photo Manager server first extracts the name of the specified photograph and queries the database for the date of the photograph. The Photo Manger queries the database for all photos taken within the selected photographs date ± x, x being the interval specified in the date-query frame. These photographs' full URL location are returned to the Hub in an XML string.

Spoken Query: In a spoken query, the Photo Manager receives a frame containing the n-best list from the retrieval recognizer representing the top n hypotheses the recognizer believes the user spoke. As an example, if the user spoke "Show me John Doe's photos of Julia at Disney World from December 2005.", and entry in the n-best list would be in the form:

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

If any metadata information, such as annotation speaker or a date or date range, have been specified in any entry of the n-best list, that information is extracted. The query terms are also extracted and parsed into words. Every word which appears in the query terms which is not a stop word and appears at least 3 times or is in the top 3 n-best results, as well as the number of appearances of the word in the query terms is held. These numbers were chosen so that only words which appear often in an n-best list, or words which have a high probability of being correct are used to query the database for photographs.

Every word still held is used to query the database for photographs. The format of all the tables discussed are detailed further in section 3.4. The Annotindex table is used to identify the ID number of the associated word. The Annotwordpic table is then used to identify the ID numbers of any photograph which contains the word in it's n-best list of annotation hypotheses. The Pictures table is then used to find the name and information of any photograph which both contain the word in its n-best list and pass any metadata predicates. An example of the SQL command to retrieve photographs for which match a given word is shown in Figure 12. This information is used to calculate a photographs TF-IDF score for a given word (Section 2.2). These values for each photograph are then summed and weighted to get the total TF-IDF score for a photograph.

```
SELECT p.filename, p.speaker, p.set, p.id, p.picdate, a.annotation,
ai.pic_freq, awp.count FROM pictures as p, annotations as a, annotindex as ai,
annotwordpic as awp WHERE ai.id = awp.wordid AND awp.picid = p.id AND p.id =
a.picid AND a.n_best_pos = 0 AND ai.word = 'queryword' AND p.speaker =
'john_doe'
```

Figure 12: A generic SQL command to obtain photos which have a query word in their annotation.

The photographs returned from the spoken query are then sorted twice to organize the photos and display higher ranked photos higher in the results list. First the photographs are sorted by the speaker of the annotation, the order of the speakers being determined by the highest TF-IDF score for a photograph spoken by that user. Then within annotation speaker group, the photographs are sorted by TF-IDF score from highest to lowest. This ensures that the photographs ranked highly either had many appearances of words in their annotation which also appeared in the query terms, had appearances of words which appear in their annotation which also appeared in the query terms but do not appear in other annotations, or both. These photographs' locations, with pertinent metadata information, along with the n-best hypotheses from the retrieval recognizer are stored in XML and sent to the web GUI via the Hub and the GUI manager.

3.4 Postgres Database

A Postgres database holds all pertinent information about photographs, including their location, any metadata associated with the photographs, and an annotations n-best list of hypotheses from the recognizer. The n-best list is also parsed and indexed by word so that photographs can be found quickly from query terms, and so the TF-IDF scores associated with the photographs can be quickly and easily calculated.

3.4.1 Pictures Table

The pictures table contains metadata information about the photograph. Specifically, the owner of the photograph, the set of photographs it appears in, the filename of the photograph, the date the photograph was taken, a unique id, and a unique tag made from [speakername]_set[set number]_[filename] are stored.

3.4.2 Annotations Table

The annotations table holds all the n-best lists associated with an annotation associated with a picture. Each entry of the n-best list is stored in full, along with its position in the n-best list. It is referenced to a specific photograph with a picid field, which associates an photographs id from the pictures table.

3.4.3 Annotindex Table

The annotindex table stores any word which appeared in any n-best list result from any photographs annotation in the system. Specifically a word is stored, along with a unique id, and the number of photographs in which the word appears at least once in the photographs n-best list result. This last field is stored so that the TF-IDF score of a query for a specific word can be quickly calculated (Discussed in section 2.2).

3.4.4 Annotwordpic Table

The annotwordpic table relates a photograph in the pictures table to a word in the annotindex table. Specifically an entry in the annotwordpic table has a photographs id from the picture table, a word id from the annotindex table, and the number of times that the specified word appears in the n-best list of the specified photograph. This table allows for quick lookup of photographs which contain a specified word. The number of appearances of a word in an n-best list is stored for calculation of a TF-IDF score.

3.4.5 Transcriptions

Some users annotations have been hand-transcribed for later use. Although currently not used in the system, the tables to store the transcription have been created and are populated. The tables follow the same format of the annotation tables, storing the transcription, words which appear in a transcription, and a table linking a word to a photograph's transcription.

All of these tables are shared between the annotation systems (section 3.1 and 3.2) and the retrieval systems (3.3). Through the storage and indexing of a photograph's annotation as well as important metadata about both the photograph and the database, the retrieval system is able to quickly return a sorted set of results to the speaker.

Chapter 4 Experiments

This chapter explains and discusses the experiments done on the retrieval system to test its effectiveness. To test the photo retrieval system, 594 verbally annotated photographs were collected from ten different users to populate the database using the phone-based annotation system (section 3.1). 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 in (section 3.2.1), annotations were collected on a Nokia N80 mobile phone at a sampling rate of 8kHz and sent in batch to the system.

4.1 Experimental Conditions

To experimentally test the retrieval capabilities of the system, seven of the nine users who provided their own annotated photographs were used as test subjects. Each subject was requested to use the web-based photo retrieval system 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. No date information was used for the queries because of the current unreliability of digital cameras to store the correct date, if it is stored at all. 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 mis-recognitions. After five successive failures, the subject was asked to move onto a new photo.

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 13. 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.

It was 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, it was expected that there would typically be agreement between the annotator and the test subjects on the vocabulary items used to describe the photos. It was also expected that the generic nature of the photographs would result in annotation vocabulary items that were likely part of the recognizer's vocabulary. It was 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.

It was 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 the name of the owners/annotators of the

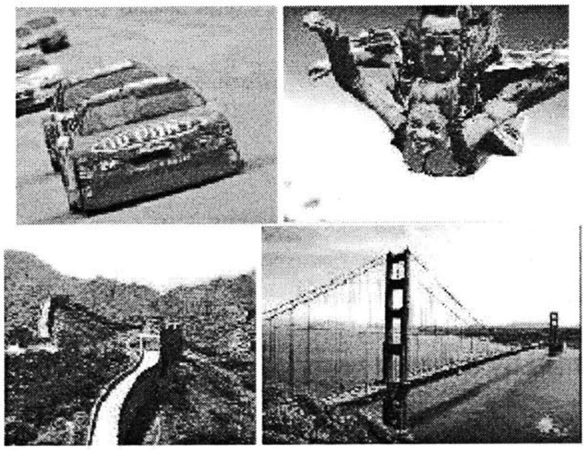


Figure 13: Examples of set A photos; generic photos all test subjects retrieved

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 the Set C test photographs.

4.2 Retrieval Results

Table 1 summarizes the results of the 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.

# of Queries	Set	% of Photos Retrieved	
		As Top Choice	In Top Ten
On First Query	A	63.3	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. Set A are the generic photos, set B are the photos annotated by other users, and set C are photos annotated by the the test subject.

In the table, the results mostly matched the 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 originally hypothesized. 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 I anticipated. I 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.

While these retrieval numbers do not guarantee a usable retrieval system currently, especially on a mobile device where only one photo at a time can be shown, as a first effort into this domain I believe the retrieval numbers are positive. As improvements to the recognizers are made (chapter 5) the errors with recognition will be dramatically reduced and better indexing techniques can be used. Also, users who repeatedly use the system both for annotation and retrieval can get a better understanding of how the recognizers work and can use this knowledge to improve their annotations and queries, thus improving retrieval results.

Chapter 5 Summary and Future Work

This research was an effort to create a system which allows users the ability to annotate and retrieve their photographs through speech. Speech recognizers were developed for the specific tasks of photograph annotation and retrieval. Also, a TF-IDF scoring system was created for ranking photographs returned from spoken queries. Through developing the systems and testing their usability, I have shown that there is a potential for users to easily annotate and retrieve their photos, a task which many people often find difficult. Also, by using speech as an input modality, the retrieval system could be extended to use mobile devices where key based input is often difficult. As these systems improve, there are several areas where this research could be expanded along with features that can be added to the results to allow users to retrieve photos more efficiently.

The current system does not use the probability associated with an n-best result to aid with retrieval. This could be stored along with the word index to give a more accurate TF-IDF score so that words which the recognizer gives a higher probability will result in a higher TF-IDF score relative to other words. This could potentially give the user a better ordering of retrieved photographs.

The current system uses an n-best list of annotations and retrieval queries to fetch photographs. Another potential approach is to store annotations as phonemes, creating a reverse index on n-gram phoneme "words". Because of the wide vocabulary of photo annotations, many words in our original system could be stored incorrectly. The n-gram phoneme search would solve this problem by storing annotations based on the sounds made, and not the words hypothesized by a recognizer.

Acoustic matching, developed by A. Park [9], could also be used to retrieve photographs with similar sounding annotations. In this scheme, photos can be related by a distance metric representing how similarity between their annotation's acoustic recording. This again could be used to aid in retrieval of photos which have been tagged with out of vocabulary words because if one photograph with that word can be found, a "similar sounding" search could be performed to retrieve other photographs. Unfortunately this can't be used for spoken queries both because the system would not scale and because users' spoken queries are often very different from the spoken annotations they are searching for.

Summarization of retrieval results would allow users to quickly understand some of the important characteristics of their search. For instance, the system could automatically count up how many photographs correspond to speakers, or how many photos occur in each month of the year. This information could be used to quickly gauge the success of a search, without having to look through all the photos returned. Especially if this system is moved to a mobile phone where only one photo can be seen at a time, this will greatly improve the user experience retrieving photos.

Another enhancement to the user retrieval experience would be to allow users to narrow down their previous result set with another query. For instance, a user could say "Show me John Doe's photos of Julia." Which may results in a large set of photographs. The user could then narrow down that set of photos with a second query "Narrow down those photos to those from Disney World." Similar to summarization, this would allow users to create better searches without having to sift through an entire result set, and important feature on phones where only one photograph can be seen at a time.

Lastly, both the annotation and retrieval recognizers could both be improved through data collection and the transcription of data. While our current system does run from some data collected from a web hosting site, the vocabulary used when speaking an annotation could be very different from the vocabulary used when typing. It is important to get users to speak naturally into the system to get a better understanding of what words are used and use that information to create a better and more accurate recognizer.

Acknowledgements

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

References

[1] J. Makoul, 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 Transaction 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 *Proceedings Eurospeech*, pp. 1229-1302, Aalborg, Denmark, September 2001.

[4] J. Chen, T. Tan, P. Mulhern, M. Kankanhalli, "An Improved Method for Image Retrieval Using Speech Annotation." Taiwan: *Proceedings of the 9th International Conference on Multi-Media Modeling*, pp.15-32, January 2003

[5] R. Srihari, "Use of Multimedia Input in Automated Image Annotation and Content-Based Retrieval." *Proceedings of SPIE*, Volume 2420, pp. 249-260, February 1995

[6] 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

[7] G. Salton, C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval." *Technical Report: TR87-881* Cornell University. 1987

[8] S. Stephanie, *et al*, " Galaxy-II: A Reference Architecture for Conversational System Development." *Proceedings of the 5th International Conference on Spoken Language Processing*, vol. 3, pp. 931-934, November-December 1998

[9] A. Park, "Applications to Word Acquisition and Speaker Segmentation." PhD Thesis. Massachusetts Institute of Technology. 2007