Web-based Semi-Automatic Annotation for Digital Photographs Using Speech and Graphical User Interface

Siti Azura Ramlan  
Graphic and Multimedia Department, FSKSM, 81310 UTM  
+6012-7922213  
ct_zura2007@yahoo.com

Sharir Daut  
Graphic and Multimedia Department, FSKSM, 81310 UTM  
+60133632215  
sharir2@yahoo.com

Nor Azman Ismail  
Graphic and Multimedia Department, FSKSM, 81310 UTM  
+6075532068  
azman@fsksm.utm.my

ABSTRACT
Annotation process is important because it has a direct impact in the way in which photo retrieval is performed in locating the photos from the photo database. However, users want an efficient approach to access or retrieve photos based on metadata of images. So, annotation process would be vital and operational going to archive high user performance in photo retrieval process. In this paper we describe our ongoing work on web-based semi-automatic annotation for digital photo using speech/graphical user interface (S/GUI) which allows user to use speech other than traditional annotation methodology by using mouse and keyboard. We applied a semi-automatic annotation process based on semantic keyword which divided by four categories called as a photo metadata: what, who, when and where.

Categories and Subject Descriptors  
D.2.11 [Software Engineering]: Software architectures; H.4.m [Information System Applications]: Miscellaneous

General Terms  
Management, Performance, Design, Human Factors.

Keywords  
Image Annotation, Image Indexing, Multimodal Interaction, Human-Computer Interaction, Speech Recognition.

1. INTRODUCTION
Nowadays, popularity of digital camera technology is increasing from year to year. We also can see that the sales of photo paper increase consequence by use of these technologies. With the good quality (large mega pixels) of digital photo captured by current technology of mobile phone with digital camera contributed to people have their digital photographs. All these factors encourage the investigation of web-based annotation for digital photo using speech/graphical user interface. Hence, we can see that people are now more comfortable to keep their photographs in digital form.

This situation has motivated research in the architecture of digital photo management system. Kerry Rodden has done his research in 1999 about how people manage their photographs. Rodden found that people would like to have systematic photo management, easy in browsing but they do not make the effort to do it [1]. After a few years, Rodden and Kenneth analyze one system for photo management called Shoebox [2]. Shoebox is a software package for organizing, annotating, indexing, searching and browsing of digital images [2]. In personal photo management annotation process could affect the effectiveness of photo retrieval task. Annotation can be the process of photo labeling by using any keyword to demonstrate the actual meaning of the photo. The label is determined to assist photo organizing and photo retrieval process.

We propose multimodal architecture specific in web development for photo annotation process by using speech input modality. Combination traditional method and speech will give many advantages in this system rather than text only. Then speech is more efficient, does not require a keyboard because majority of people can speak faster than write or type [3].

2. RELATED WORK
Previous work in photo annotation has largely focused on photo management organizer. Myphotos[4] is a system for home photo management but this system did not have specific feature for annotation processing. On the other hand, Shoebox[5] is a system for the photo management of personal photo that included with spoken annotation facilities. While, Fotofiti[6] and EasyAlbum [7] is a system to help users in annotation photo. Currently from our survey, semi-automatic photo annotation is better than manual annotation in terms of efficiency and better than automatic annotation in terms of accuracy [8].

This paper inspire from J.Chen , T.Tan and P.Mulhem project. They introduce a methodology for image indexing using speech annotation technique [9] and following that their next paper [10] are specifically for image retrieval. Meanwhile, B.Suh and B.Bederson introduced an approach to semi-automatically annotation using timestamps and use a torso based human recognition technique [11].

In the next section, we explain an overview of on going prototype and how the system work.
3. SYSTEM OVERVIEW

3.1 System Architecture

Figure 1 shows the system architecture of web-based photo annotation using speech/graphical user interface. The system can comprised of four components. The major components are Development Tools specific to develop the system, Clients, Server included database and Speech Engine Services.

![Figure 1: System Architecture](image)

The combination of all sub-components constitutes the best of integration for each major of components. Client describe as a part where user interacts with the system by using HTML, JavaScript and SALT code. Function of clients is as users’ interface that allows user to communicate with the system through speech dialogue. Processing of speech command obtain in speech engine that consist of speech recognition and speech synthesis. All the process has executed in the Server whereas a place to store the grammar and metadata including all photo that we assume already in the database.

The scope of this research limited for personal photographs where digital photo are already uploaded to the Server. It just focused on annotation part but we are considering the effects in photo management and particularly in photo retrieval process. The annotation processing goes up to achieve highest usability in user performance perspective. Flow of annotation process indicates in the next sub-section.

3.2 Annotation Process

What the keywords present by user have used during annotation process called as metadata of images. For example, user creates some word on the back of photo. That action we can call as annotation process. The important things in annotation process is a metadata of image where consider about what appropriate keyword to indicate the content of photos. From that, there are many researches have simplify the image annotation process using four categories as a metadata. They consist of who (indicate people on the photo), what (indicate as a subject), when (commonly use timestamp) and where (to indicate where the photo taken). Related with metadata, M. Naaman et al [12] developed automatically generate an abundance of photo related metadata using off-the-shelf and web-based data sources based on location information on digital photographs. They were using are location, time of day, light status, weather status and temperature, and season as a metadata.

![Figure 2: Annotation process](image)

Table 1: Annotation Techniques

<table>
<thead>
<tr>
<th>Human Effort</th>
<th>Machine Assistance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>Save annotations to database</td>
</tr>
<tr>
<td>Semi-Automatic</td>
<td>Parse keyword-oriented concepts and extract semantic information.</td>
</tr>
<tr>
<td>Automatic</td>
<td>Add structural annotations using GPS, context, or recognition technology.</td>
</tr>
</tbody>
</table>

Annotation technique consists of three categories, which are manual, semi-automated, and automated. Table 1 are the advantages and disadvantages for each technique that provided by J.Kustanowitz and B.Shneiderman [13]. The discussion of this paper focused on semi-automated technique, which is with users desired and it is a friendly system. This is because semi-automated technique can be a freestyle process and fulfill users desired.

Figure 2 indicates the flow of annotation processing where we assume that the photo already have in the database. Start the process with photo browsing but the list of photo position is not in order. Then, photo and annotate form are displayed based on user selection from the list of photo. Continue with an assigning signal for speech, user put out the words and speech recognition engine will be change the voice into text format. This process shows in figure 3 where it is a multimodal interface can be controlled input output by multiple modes or channels of communication such as speech with mouse, speech and keyboard or both of them. Speech will start when user click at activate button for speech. If errors occur along the speech process, users need to repeat the first step (assign signal). In the mean time, if there are no errors, data (input from speech) in voice will convert into text format and send into the database. When filling form complete, all data are forwarded.
to the database and saved. Lastly, the photo and its metadata are displayed after storage of data completed.

### 3.3 Selective Metadata

Kuala Lumpur, Cats, New born and January 23 are the examples of keywords that users use to labeling their photos. All words are metadata of images.

![User Interface for Annotation](image)

Refer to the figure 3, the multimodal interface contains five rows of field including description and all categories can be indicating in the following table.

<table>
<thead>
<tr>
<th>Category</th>
<th>Representative</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td>Time and date</td>
</tr>
<tr>
<td>Who</td>
<td>Subject</td>
</tr>
<tr>
<td>What</td>
<td>Event</td>
</tr>
<tr>
<td>Where</td>
<td>Location or place</td>
</tr>
</tbody>
</table>

### 3.4 Speech Environment

We found that, Windows Speech Recognition in Windows Vista enables users to interact with their computers by voice. Users can dictate documents mainstream applications and use voice commands to control the operating system. Dragon Naturally Speaking, a product of Nuance also provides users the power to create documents, reports, emails and more all by speaking.

In the section 1, we have represented that people commonly speech faster than write or typing. Dialogue system builds up from a conversation between two or more side of users. Here, we are focused on interaction between human and computer (can be Pocket PC).

Microsoft Speech Application SDK gives an easy way to develop speech web applications. Then, all speech web application is based on SALT tag where each speech control corresponds to a SALT tag. For this prototype, we propose using SALT as a speech interface markup language specific to write speech interface for the web application with voice input output in terms of speech recognition and speech synthesis. SALT containing of four primary XML tags consists of prompt, listen, DTMF (dual-tone multifrequency) and smex tag.

```xml
<salt:grammar xmlns="http://www.w3.org/2001/06/grammar">
  <grammar root="root">
    <rule id="root">
      <item repeat="0-1">from </item>
      <ruleref name="#cities" /> </rule>
    </grammar>
    <one-of>
      <item> Cambridge </item>
      <item> Seattle </item>
      <item> London </item>
    </one-of> </grammar>
  </salt:grammar>
```

Figure 4: Inline grammar

Figure 5: Reference grammar

Grammars are a set of nonterminal and terminal symbols, and a set of production rules. The grammar element is used to specify grammars, either inline or referenced. Figure 4 shows the code of SALT include the grammar in one code instead of figure 5, the grammar represent as a file and call the grammar file in the “src” element.

### 4. FUTURE WORK

In future work, we will continue with high fidelity related prototype and analyze the users’ performance of the system.

### 5. ACKNOWLEDGMENTS

Thanks to GMM, FSKSM, through out the progress of this research.

### 6. REFERENCES


