GROUP ANNOTATION FOR DIGITAL PHOTOS USING MULTIMODAL INTERACTION

S. A. Ramlan, S. Daut, N. A. Ismail
Department of Computer Graphics and Multimedia,
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia
81310, Johor, Malaysia
email : ct_zura2007@yahoo.com

ABSTRACT

Making a note for each digital photo one by one is not efficient when the photo collection is thousands of units. Therefore, users will want an efficient approach to retrieve photos based on metadata of their digital photos. So, annotation process would play a vital role in the retrieval process. In this paper, we describe our growing work on a group of photo annotations for personal digital photos by using multimodal interaction. The developing prototype allows us to annotate their photo in World Wide Web (WWW) environment. Multimodal interaction will give users an alternative way to interact with the system. We apply a human computer speech dialogue in this system based on semantic keyword which is divided by four categories of digital photo metadata: what, who, when and where. The proposed prototype will be developed by using web programming scripting language containing Javascript, XHTML, PHP and XML Language Tag (Speech Application Language Tags (SALT)).

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

1 INTRODUCTION

People now have thousands of digital photo collection as a consequence of digital camera popularity and mobile phone with camera technology. We can also see that the sales of photo printing paper have risen due to the rapid growth of these technologies. This situation effected users to keep a variety of photos for their own reasons. Photos are valuable things because they are powerful in expressing the implicit storytelling. They could also be recognized as a memorial for a particular event. Useful annotation process method for digital photo collection is necessary so that photo management becomes easier especially when users want to search or browse through their photos. This situation has motivated research in the area of digital photo management system. In 1999, Kerry Rodden found that people would like to have a systematic photo management and easy access in browsing, but they do not make the effort to do it [1]. A few years later, Rodden and Kenneth analyzed a system for photo management that could organize, annotate, indexes, search and browse digital images called Shoebox [2]. The significant part in photo management is the annotation process. The process of labeling the semantic content of photos (or objects in photos) with a set of keywords or semantic information can be known as photo annotation. Photo annotation aims to create semantically meaningful labels and associate users with their photo [3]. The information collected from the annotation process plays a crucial role for photo management, especially for browsing, recalling, finding and searching photos. Photo metadata in the annotation process is the data that users use to tell about related photo. In our observation (reported elsewhere), photo collection has a probability to share the same metadata in describing the photo. For instance, two photos taken at the same event will share the metadata of “what” in terms of name of events (example: birthday party). M. Naaman et al [4] did their research concerning the contextual metadata that assemble a photograph automatically utilizes time and location metadata. A paper survey by S.J. Cunningham explained that people are more likely to describe their photo by a set of photo rather than a single photo [5]. Therefore, users need the effective way to send data for multiple photos into the database just by one click.

We propose an annotation process in WWW environment for a group of photo annotations by using multimode input. We also implement human computer speech dialogue in our prototype. The combination of traditional method and speech will give incentives for users in usability of this system rather than the traditional way only. This is because speech is more efficient, does not require a

ISSN 1858-1633 @2008 ICTS
keyboard and majority of people can speak faster than write/type [6]. All of these factors have encouraged the investigation of web-based annotation for a group of digital photos instead of annotating each photo one by one that requires user to spend a long time to complete the annotation process of all photos.

2 RELATED WORK

There are now a number of research led-system for managing digital photo collection, including Fotofiti, MyPhoto and Shoebox. MyPhoto [7] is a system for home photo management and processing that includes an auto grouping method for easier browsing, searching and organizing. Fotofiti [8] focuses on personal photo organizer with automatic image annotation, event management and social network integration on the web. Although these systems do not employ speech recognition to provide audio annotations, they do place great attention in making annotations easier for users. On the other hand, Shoebox [9] provides annotations using speech recognition for personal digital photograph collections, which is similar to our work. It also provides an application for image retrieval both semantically similar and visually similar by voice.

EasyAlbum [10] introduced a method of annotating photograph based on face clustering and re-ranking. Cluster annotations allow annotating photos by group instead of one by one. Meanwhile the re-rank intention is to arrange similar photos close to each other so users will find it easier to label them together. B. Suh et al [11] discussed on annotating photos based on hierarchical events clustering or multiple levels of event groups and torso based human identification. Torso based human identification was chosen due to low accuracy in human face recognition. Nevertheless, this system fails to work properly for photos contain people wearing clothes with pictures or uniforms. ImageGrouper [12] incorporated a concept called Query-by-Groups so users can construct query for groups of images. These groups of images are created by dragging the images on the interface. This operation allows users effectively annotate and organize a large number of photos. In indexing photos, J. Chen et al [13] define four fields which are people, place, date and event that can be implemented in annotating photos. But they stated using speech recognition tools by Nuance Dragon Naturally Speaking to perform the speech to text role. It may take a while for users to learn the speech tool used in this interface. Meanwhile, K. Christian et. al [14] have done their study about a comparison of voice controlled and mouse controlled web browsing. Their study shows that navigation by voice control raised the performance for the certain task such map navigation. While, L. Karl et. al [15] in their journal shows that performance in timing for all tasks were significantly faster when using speech to activate commands.

This paper is inspired from B.Suh and B.Bederson project. They introduced a methodology for photo algorithms for generating meaningful photo groups in two techniques which are hierarchical event clustering and clothing based person recognition [16]. The scope of this research is limited for personal photographs where the digital photos are already uploaded to the Server. Our focus in this research is on the group annotation part, but we are considering the effects in photo management and particularly in the photo retrieval process. The next section will discuss the propose annotation system and low fidelity prototype. We cover an overview of the ongoing prototype and the annotation process.

3 SYSTEM DESCRIPTION

The popularity of metadata collection has increased in web-based environment. Web application for photo management with rich metadata through tagging has become popular with photo management application such as Flickr and Fotofiti [17]. Annotation in Flickr serves for personal and social purpose, increasing incentives for tagging and resulting in a relatively high number of annotations [18]. Among of the annotation is tagging approach which is using short keywords where keywords are related with the object (photos) [17]. Annotation includes speech interaction will take more advantages of advances in speech technology. Show&Tell [19] is one of the research that combines advances speech recognition, natural language processing and image understanding. Show&Tell application also suits well for medical images other than NASA photo and consumer photos. The system description for this prototype is particularly for personal photo collector.
Figure 1 shows the architecture of the web-based annotation system including the development tools that we use to develop the prototype system. The system architecture consists of two parts, which are server and client. Web Server is a place to store all the data, programs, grammars, prompt and annotation engine. The client side is the user's interaction part consists of multimodal web browser, input/output devices, SAPI (Speech Application Programming Interface) that allow the use of speech recognition and speech synthesis within windows application. At the client side, we specify a desktop PC running a version of Microsoft Internet Explorer that support speech application, and it will interpret with XHTML, SALT (Speech Application Language Tags) and Javascript. The prototype system will be developed on web-based platform. Users have two selections in using their own style of interaction either to use Graphical User Interface (GUI) environment or Speech/Graphical User Interface (S/GUI) environment to annotate the photos based on four categories of photo metadata (what, who, where and when). With the selection of modalities inputs, users can apply any input that they want. For this prototype, users could annotate multiple photos at once. Therefore, users will feel more flexible and comfortable in terms of using group annotation or single annotation.

3.1 Annotation Process

The Annotation process describes the flow of photos tagging using multimodal interaction. Users have two choices of an interaction way either using group annotation or single annotation. We can see in a learning situation in which teacher will give any information at once to a group of students. For example, “Michele, Johanna and Ruby, please find out the solution for question 5 on page 23”.

Related with metadata research, M.Naaman et al [20] developed an automatically generate an abundance of photo related metadata using off-the-shelf and web-based data sources based on location information on digital photographs. Location, time of day, light status, weather status and temperature, and season were use as a metadata. This prototype is towards a flexible application enabling the use of speech as an alternative way to communicate with the system. Beside that, users have the liberty choosing the style of annotation in either group annotation or single annotation. Even so, in this part we introduced two choices of interaction that shows in Figure 2 and Figure 3. There are two flowcharts comprised of the traditional way and the speech enabled annotation process.
Figure 2: Annotation process
(Mouse+Speech)

Refer to figure 2, the process starts with browsing by saying the photo with “select” before the number. For example “select 2”, then related photo will be checked. Then, user utilizes “finish” command to assign complete selecting and the process move to part of the annotation process (right hand side of content on figure 4). To enter the input of metadata, user needs to click the microphone button to activate speech interaction mode (refer figure 4). After finished with no error, user will submit all data into database. Meanwhile, the flow of the annotation process without speech interaction is on figure 3. It starts with photo browsing and ends the process when the data saving is completed and no error along the annotation process.

### 3.2 Dialogue Management

There are many researches related with a dialogue system in various applications. Dialogue design is important and should be considered towards user’s satisfaction. Research into multimodal interfaces indicates a distinct advantage to using speech as an input mechanism [14]. The combination of the advantages of graphical, spoken and direct manipulated interface could ease the use of the system [21]. But, speech input is not likely to replace other modes of input, but seems to have a useful place along with other types of input in a multimodal interface.

In dialogue management, we are also considering expert users where they can use flexible textbox on the top of content part to enter the input of photos metadata. They need to use “annotate” following with a category of metadata. “Annotate” command is presented as a signal for annotating photos. Then, say the input to the system and submit all of data using “data submit” command. Hence, all of the command words consider with the suitability of interface design and level of users' expertness in using computer.

### 4 CONCLUSION AND FUTURE WORK

This paper presents the propose prototype system in web-based annotation for digital photo collection. Currently, the low fidelity prototype design is completed. From the low fidelity prototype, we can see the real incoming situation of the annotation process. In this paper, we described the system architecture, later with annotation process flow included a traditional way and combination of speech and mouse and dialogue management for an annotation process.

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

### 5 ACKNOWLEDGMENT

Thanks to Department of Computer Graphic and Multimedia, Faculty of Science Computer and
Information System, Universiti Teknologi Malaysia through out the progress of this research.

REFERENCE