Preprint – Final version appears in Proceedings of the International Conference on Computer and Communication Engineering 2008
May 13-15, 2008 Kuala Lumpur, Malaysia

Enabling Multimodal Interaction in Web-Based Personal Digital Photo Browsing

N.A. Ismail[1] and E. A. O'Brien[2]

[1] *Department of Computer Graphics and Multimedia, Universiti Teknologi Malaysia, 81300 Skudai, Malaysia*
azman@utm.my

[2] *Research School of Informatics, Loughborough University, LE11 3TU, United Kingdom*
A.O-brien@lboro.ac.uk

Abstract

Retrieval process of both digital photos and physical photos has not been easy, especially when the collections grow into thousands. In this paper, we describe an interactive web-based photo retrieval system that enables personal digital photo users to accomplish photo browsing by using multimodal interaction. This system not only enables users to use mouse clicks input modalities but also speech input modality to browse their personal digital photos in the World Wide Web (WWW) environment. The prototype system and it architecture utilize web technology which was build using web programming scripting (JavaScript, XHTML, ASP, XML based markup language) and image database in order to achieve its objective. All prototype programs and data files including the user's photo repository, profiles, dialogues, grammars, prompt, and retrieval engine are stored and located in the web server. Our approach also consists of human-computer speech dialogue based on photo browsing of image content by four main categories (Who? What? When? and Where?). Our user study with 20 digital photo users showed that the participants reacted positively to their experience with the system interactions.

I. Introduction

In everyday life, people already have large collections of printed personal photos and the new digital photo technology helps collections grow further [1]. This technology has resulted in more people having large personal collections of digital photos and sharing them with others, especially with their family and friends. Web based photo galleries are one of the methods of sharing that allow photo users to publish a collection of digital photos online in a centralized and organized way [2].

Nowadays, there are currently a numbers of digital photo systems that have different retrieval approaches and have commercial production quality, experimental systems and freeware packages. Among them are Apple iPhoto [3], Ulead iMira [4], Adobe Photoshop Album [5], Personal Digital Historian [6], FotoFile [7], AT&T Shoebox [8], PhotoTOC [9], PhotoFinder [10], Flickr [11] and various packages bundled with digital cameras. Some of these systems provide browsing, free text searching and even a range of limited visual content based retrieval.

Advances in automatic speech recognition engine and multimodal web browser have make multimodal user interface which support speech based modes of interaction are possible in World Wide Web (WWW) environment. Recent studies indicate that there may be advantages to have an additional input channel based on speech input placing along with other types of input modalities in a multimodal interface [12, 13].

All these factors encourage the investigation of integrating multimodal interaction styles for browsing mode into the web-based personal digital photo retrieval system.

II. System Description

Our system (FlexPhoReS) differs from previous work in the area of system environment and retrieval strategies. The prototype system is based on WWW environment and could allow users to use multimodal interaction which refer to the style of interaction that enable users to use either mouse click (Graphical User Interface (GUI) environment) or "mouse tap and talk" input modalities (Speech/Graphical User Interface (S/GUI) environment) to browse their digital photo via a set of user-oriented categories 4W's (Who? What?

When? and Where?). Therefore the user could select the input methods that best suits their browsing tasks. This was believed could give more flexibility to the personal digital photo collector. Figure 1 shows the schematic diagram of the prototype system that have the abilities to browse personal digital photos by event (What?), by place (Where?), by people/subject (Who?), and by time (When?) from user's photo repositories web database using multimodal interaction.

The prototype system also allows user to control or navigate through the system using multimodal interaction. For example to logout from system, go to the main page, retrieve system help, and go to next page and previous pages.

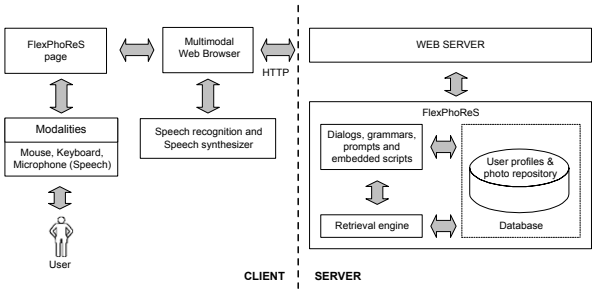

Figure 1. Architecture of the FlexPhoReS

The prototype system and it architecture utilize web technology which was build using web programming scripting including JavaScript, XHTML, ASP, XML based markup language – Speech Application Language Tags (SALT) and image database. All prototype programs and data files including the user's photo repository, profiles, dialogues, grammars, prompt, and retrieval engine are stored and located in the web server. The client machines run the web browser and the server machine runs the web server (Figure 1). Microsoft Internet Information Services (IIS) web server was used to deploy FlexPhoReS prototype system in the WWW environment. For client perspective, internet explorer with speech add-on is used as multimodal web browser which allows users to run speech technologies along with keyboard and mouse for multimodal interaction.

Through multimodal interaction, users could select the interaction modes that best suits their requirements to perform browsing task. For "mouse tap and talk" input modalities the user can tap to activate the microphone and speak specific words. There are 3 different microphone buttons with different embedded functions for speech interaction. Figure 2 is a snapshot of FlexPhoReS user interface.

Figure 2. FlexPhoReS user interface

The first microphone button refers to photo browsing and application control functions. The second microphone button refers to searching by keywords function and the third microphone button refers to searching by visual similarity function. However the second and the third microphone functions of the prototype recently reported elsewhere.

Browsing Commands and Dialog Management

In FlexPhoReS, photo browsing was based on four different categories of browsing. Users could browse photos by clicking on the categories of Event, Place, Subject/People and Time. Each category of photos was already associated by hypertext with retrieval words that link to the related user's photo collection. Users simply choose to browse their photo categories by clicking on the retrieval words hyperlink and FlexPhoReS displays the set of photos (result set) based on the chosen hyperlink word. When using "mouse tap and talk" input modalities, users have to click the specific (blue) microphone to invoke Browse

mode and identify the appropriate browsing categories by using speech. If the data is not recognized, the system will prompt an error message through speech output and ask the user to re enter the input. This process will continue until the user speaks the recognized input data. 'What can I say?' hyperlink is a medium for the user to know what they can speak if they tap the selected microphone. The hyperlink will also pop up if the user taps the microphone and asks 'what can I say?'

To browse the photos, there are four possibilities: Event, Place, Subject/People and Time. Speech recognition gets more difficult when the application grammar and vocabularies are large or have many similar sounding words [14, 15]. At the recognition stage, due to performance limitations, speech recognition is also unstable when recognizing too many combinations of vocabularies or long words input. The users therefore need a simple word to invoke the correct browse category in order to avoid any grammar collision with other photo browsing retrieval categories [16]. For example, if the user desires to search for photos of York, he/she has to say the terms "York" and also the category term "Place". In the same way, if the user wishes to retrieve photos of snowing, the user has to say the word "snowing" and the category term "Event".

III. USER EVALUATION OF THE PROTOTYPE

Twenty digital photo volunteers took part in the final evaluation. All of the participants had experience retrieving personal digital photo by browsing. The participants who took part in this evaluation were digital photo users recruited randomly from various backgrounds at Loughborough, United Kingdom. The purpose was to examine their browsing performance using multimodal interaction and the acceptability of the prototype system interaction. A set of tasks was devised for the study. User interaction with the interface was recorded by screen and audio recording software which provided a clear picture if a user was successful or not to complete photo browsing as well as time taken to complete the browsing tasks.

On average participants needed less time to complete photo browsing tasks when they used "mouse tap and talk" input modalities compared with mouse click input modalities. Participants took 0.56 minutes to complete browsing tasks with "mouse tap and talk" input modalities whereas they took 0.84 minutes with mouse clicks input modalities. On average, the reduction in search by browsing performance time due to using "mouse tap and talk" input modalities was 33.3%.

Table 1 shows the means and standard deviations of the time taken to complete photo browsing tasks by all participants while Figure 3 shows the actual distribution.

TABLE 1: MEANS AND STANDARD DEVIATIONS OF TIME TAKEN TO COMPLETE PHOTO BROWSING TASKS FOR ALL PARTICIPANTS (N=20)

Description	Mean	Standard deviations
Photo browsing using mouse click input modalities	0.84	0.34
Photo Browsing using "mouse tap and talk" input modalities	0.56	0.16
Percent reduction: 33.33%		

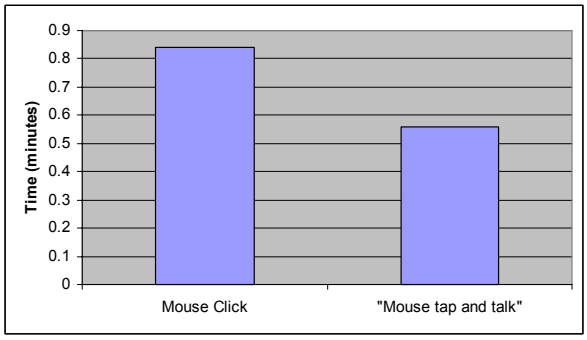

Figure 3. Average time taken to complete browsing tasks with different input modalities.

The study of acceptability of the input modalities of FlexPhoReS revealed that all of the participants agreed that mouse click input modalities (GUI environment) by themselves are suitable for photo retrieval tasks. They also agreed that "mouse tap and talk" input modalities (S/GUI environment) alone are suitable. A higher acceptability rate was given when both input modalities were considered together and the majority of participants agreed that both input modalities are complementary to each other in photos browsing. Several participants stated that both input modalities were user friendly, practical and easy to use. A number of participants noted that "mouse tap and talk" input modalities were more interesting and easier for browsing photos instead of mouse click input modalities.

IV. CONCLUSION AND FUTURE WORK

We have showed that the proposed prototype system interaction (FlexPhoReS) has yielded preliminary evidence revealing that the FlexPhoReS already provides a good basis for supporting flexibility in the web based personal digital photo retrieval interaction process. Users can execute system control and browsing commands which are embedded in the menu and hierarchical structure by saying appropriate words which may be easier than other input devices. Adding speech-based interface to support personal digital photo browsing could give users the freedom to choose and combine interactions methods to create more efficient and pleasant way of browsing their personal digital photos collection in web environment. With multimodal interaction, the weaknesses of one modality are offset by the strengths of another. Our user study with 20 digital photo users showed that the participants reacted positively to their experience with the system interactions. Our approach could empower web communities to explore their personal digital photos collection in natural mode of interaction. One of the main thrust for future work is implementing web based semi-automatic personal digital photo annotation with multimodal interaction and integrating them into one system.

ACKNOWLEDGMENT

A special vote of thanks is due to fellow research students and staff in the Research School of Informatics, at Loughborugh University, Research Assistance Siti Azura and Shahrir and staff in the Department of Computer Graphics and Multimedia. Thanks also to Universiti Teknologi Malaysia (http://www.utm.my) and Ministry of Science, Technology & The Environment, Malaysia (MOSTE) for sponsoring this research.

REFERENCES

[1] Ismail, N. A. and A. O'Brien, 2004. Towards an understanding of user needs in organising and retrieving photos from personal digital photo collections. Information Resources Management Association (IRMA) International Conference. New Orleans. Idea Group, pp. 1045-1047.

[2] House, N., M. Davis., Y. Takhteyev., M. Ames., and M. Finn. 2004. From 'what?' to 'why?': the social uses of personal photos. <http://www.sims.berkeley.edu/.../vanhouse_et_al_2004a.pdf>

[3] Apple iPhoto, 2008. <http://www.apple.com/ilife/iphoto/>.

[4] Ulead, 2004. Ulead iMira,<http://www.imira.com/>.

[5] Adobe Photoshop Album, [n.d.]. <http://www.adobe.com/products/photoshopalbum/main.html>

[6] Shen, C*., et al.*, 2003. Personal digital historian: story sharing around the table. *ACM Interactions*, **10**(2), 15-22.

[7] Kuchinsky, A., et al., 1999. *FotoFile: a consumer multimedia organization and retrieval system*. Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit. Pittsburgh. ACM, pp. 496-503.

[8] Mills, T. J., et al., 2000. *Managing photos with AT&T Shoebox*. Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. Athens. ACM, p.390.

[9] Platt, J. C., et al., 2003. *PhotoTOC: automatic clustering for browsing personal photographs*. Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing and Fourth Pacific-Rim Conference on Multimedia. Singapore. IEEE, pp. 6-10.

[10] Kang, H. and B. Shneiderman, 2000. *Visualization methods for personal photo collections: browsing and searching in the PhotoFinder*. Proc. IEEE International Conference on Multimedia and Expo (ICME2000). New York. IEEE, pp. 1539-1542.

[11] Flickr, 2008. < http://www.flickr.com/>.

[12] Deng, L., et al., 2002. Distributed speech processing in miPad's multimodal user interface. Speech and Audio Processing, *IEEE Transactions*, 10(8), 605- 619.

[13] Oviatt, S., et al., 2000. Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions. *Human Computer Interaction*, 15(4), 263-322.

[14] Shneiderman, B., 2005. *Designing the user interface*. Boston**:** Addison Wesley.

[15] Shneiderman, B., 2000. The limits of speech recognition. *Communications of the ACM*, **43**(9), 63 - 65.

[16] Hunt, A., 2004. Speech Recognition Grammar Specification Version 1.0, <http://www.w3.org/TR/speech-grammar/>