WEB-BASED PERSONAL DIGITAL PHOTO COLLECTIONS: MULTIMODAL RETRIEVAL

N.A. ISMAIL[1] AND A. O'BRIEN[2]

[1]*Department of Computer Graphics and Multimedia, Universiti Teknologi Malaysia, 81300 Skudai, Johor Bahru, Malaysia.*

[2]*Research School of Informatics, Loughborough University, LE11 3TU, United Kingdom.*

azman@utm.my and A.O-brien@lboro.ac.uk

ABSTRACT: When personal photo collections get large, retrieval of specific photos or sets of photos becomes difficult mainly due to the fairly primitive means by which they are organized. Commercial photo handling systems help but often have only elementary searching features. For many years, graphical user interfaces (GUIs) have become the user interface of choice in web applications. Moving beyond mouse and keyboard, multimodal interfaces are expected to be more transparent, flexible, efficient and powerful for human-computer interaction. 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 click 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 built using web programming scripting (JavaScript, XHTML, ASP, XML based mark-up 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 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.

KEY WORDS: *Multimodal Interaction, Speech Interface, Photo Browsing, Image Retrieval, Personal Photo and User Interface.*

1. INTRODUCTION

Digital cameras are now a common feature in everyday life. The result has been large collections of personal photos stored on home computers, which people then wish to share with family and friends [1]. Previous investigation to personal digital photos management system showed that people used the simple features effectively and reported they found their digital photos easier to organize than had been the case with non-digital photographs. The findings indicated that because people are familiar with their own photographs they usually find what they need by browsing [2,3]. These photos are often informally stored in folders and as the collection gets larger there can be problems in finding the desired images or sets. Typical practice is to store the photos on a hard disk and send by email or make use of a web based photo gallery to publish and share online in a centralized and organized way [4].

There are currently a number of digital photo systems with different approaches to retrieval, some with commercial production quality, some experimental systems and freeware packages. Among them are Apple *iPhoto [5], Ulead iMira [6], Adobe Photoshop* Album [7], Personal Digital Historian [8], FotoFile [9], AT&T Shoebox [10], PhotoTOC [11], PhotoFinder [12], Flickr [13], Google Picasa [14] 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. The effectiveness of retrieval is, however, dependent on the amount and quality of the metadata associated with the images. This is normally minimal, often simply a folder name. In many cases the images themselves will only have a numerical filename. Conventional retrieval is therefore limited to browsing.

Advances in automatic speech recognition engine and multimodal web browsers have resulted in multimodal user interfaces which support speech based modes of interaction are possible in World Wide Web (WWW) environment. Recent studies indicate that there may be advantages to having an additional input channel based on speech input placing along with other types of input modalities in a multimodal interface [15, 16].

In web applications, graphical user interfaces (GUIs) have become the user interface of choice. For many years, they have provided the user with a common look and feel, visual representations of data and direct control using mouse and keyboard input modalities as standard input devices. However GUIs only become possible to implement when computer hardware can produce accurate bitmap displays and can interactively manipulate accurate screen presentations to the users [17]. Moving beyond mouse and keyboard, multimodal interfaces are expected to be more transparent, flexible, efficient and powerful for human-computer interaction [18]. Multimodal interface technologies have been applied with some success to problems in certain domains such as personal information management (e.g. Personal Digital Assistants (PDA) applications) [15]. Numerous theoretical and empirical studies have investigated the potential of multimodal interfaces. The array of multimodal systems currently ranges from simulation and training applications to verification system security that will increasingly affect our lives [16].

Some recent studies have involved designing systems that combine either speech and pen input [19] or speech and lip movements [20]. One study has developed a multimodal interface for digital image retrieval. Käster and colleagues proposed a multimodal system that combines mouse, keyboard, speech and touch screen interface for standalone content based image retrieval [21]. The usability experiments of the study showed that users well appreciate this multimodal interface for image retrieval. It is not clear, however, that multimodal interface approaches and technologies are fully suited for web based personal digital photo retrieval applications. Personal digital photo users in general have individual differences in performance, needs and abilities in using different modes of interaction.

All these factors encourage the investigation of integrating multimodal interaction styles for browsing mode into the web-based personal digital photo retrieval system. The prototype in this study was developed to offer users flexibility in performing photo retrieval tasks.

2. 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 allows 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 4Ws (Who? What? When? and Where?). Therefore the user could select the input methods that best suit 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 has the ability 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.

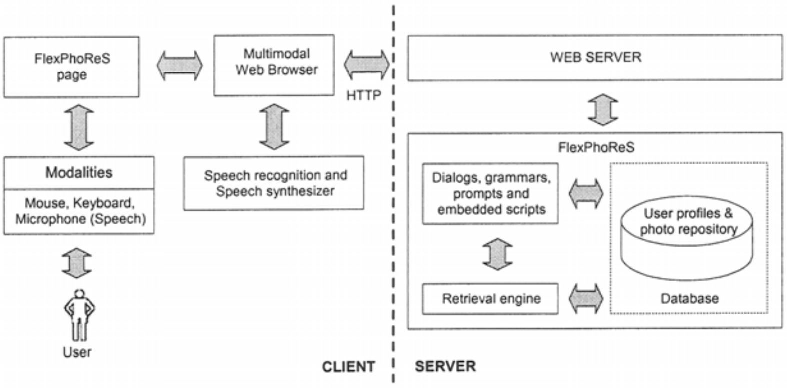
FlexPhoReS page
Multimodal Web Browser
HTTP
WEB SERVER
Modalities
Mouse, Keyboard, Microphone (Speech)
Speech recognition and Speech synthesizer
FlexPhoReS
Dialogs, grammars, prompts and embedded scripts
User profiles & photo repository
Retrieval engine
Database
User
CLIENT
SERVER

Fig. 1: Architecture of FlexPhoReS.

The prototype system and its architecture utilize web technology which was built using web programming scripting including JavaScript, XHTML, ASP, XML based mark-up 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 (Fig. 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 can 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.

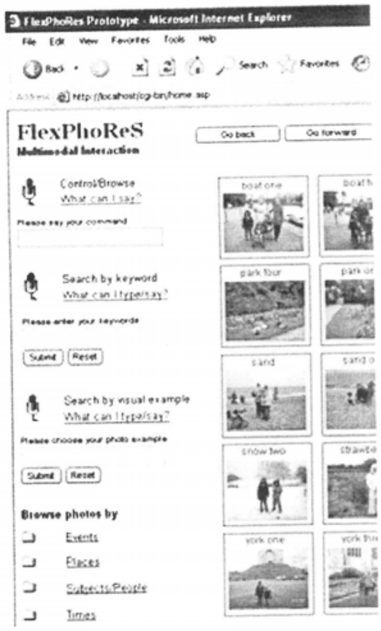

Fig. 2: FlexPhoReS user interface.

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 have been reported elsewhere [22, 23].

2.1 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 recognised input data. 'What can I say?' hyperlink is a medium for the user to know what (s)he can speak if (s)he taps 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 [24, 25]. 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 [26]. For example, if the user desires to search for photos of York, (s)he 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".

Adding speech modality as an additional interaction style provides the flexibility for users to choose the modes of interaction to retrieve their digital photos. The system is initiated when the user enters a FlexPhoReS URL in a multimodal web browser; the Web server opens the FelxPhoReS application's default page. Then the Web server sends XHTML, SALT, and JavaScript to the client machine. SALT mark-up in the pages that the Web server sends to the client can trigger the speech recognition and text-to-speech synthesis engine. For text-to-speech synthesis, the prompt element is used to specify the content of the audio output. Speech recognition, or speech-to-text, involves capturing and digitizing the sound waves. Then FlexPhoReS processes the words and compares them with the FlexPhoReS grammar (XML tag suite), which is a structured collection of words, or phrases that the FlexPhoReS recognizes and attempts to match human patterns of speech. In FlexPhoReS, the listen element is used for speech recognition. Listen element contains one or more grammar elements, which are used to specify possible user inputs. A single mode of recognition was chosen as a type of recognition scenario [27]. Single mode recognition is typically used for 'tap and talk' scenarios. In this mode, the return of a recognition result is under the control of an explicit stop call from the application. Figure 3 illustrates the single mode listen behavior of speech recognition events timeline that used in FlexPhoReS prototype system.

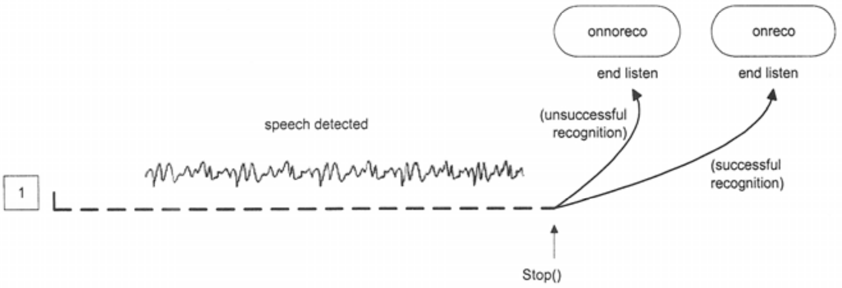

Fig. 3: Speech recognition event.

The Stop() call in action and the possible resulting listen events of "onreco" or "onnoreco". "onreco" is the event handler that is fired when the recognizer has a successful recognition result, while "onnoreco" is the event handler that fired when the recognizer was unable to return a complete recognition result.

3. 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 that 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. Because the screen and audio recording software is essentially 'invisible' it was not expected to influence users' normal tasks and searching behavior. It recorded how each participant was using the FlexPhoReS prototype.

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: 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%		

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.

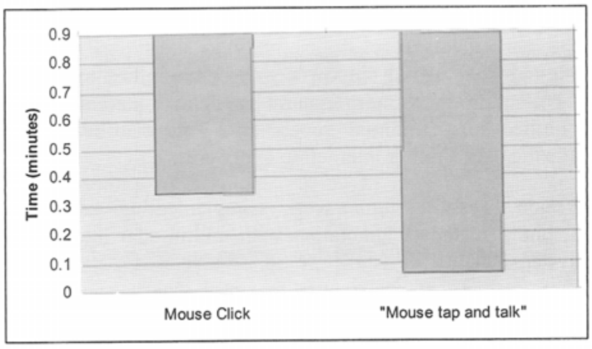

Fig. 4: Average time taken to complete browsing tasks with different input modalities.

It is clear that all of the participants had significantly improved their digital photo search performance (through photo browsing tasks) when using mouse and speech input modalities. Participants also were significantly more satisfied with mouse and speech input modalities than with mouse and keyboard input modalities. General findings from these data show that mouse and speech input modalities are faster. There are various issues to be considered here. Speaking a word or phase is generally quicker than navigating the hyperlinks on screen or typing the same text in the textbox. Although mouse and keyboard are effective input modalities in most cases, they are limited in many ways and do not provide fully natural communication between humans and computers. Speech has great potential for becoming key modality in future interactive web-based personal digital photo retrieval system.

4. CONCLUSION AND FUTURE WORK

The present study fills a gap in both the digital photo retrieval and multimodal user interface research. We have demonstrated that the proposed prototype system interaction (FlexPhoReS) advocates the use of user-oriented image retrieval 4Ws categories template (Who?, What?, When?, and Where?) for photo browsing retrieval strategies through multimodal interaction. The prototype also contributes a convenient way to manage speech elements with GUIs which include system dialogues on how to keep track of the users when the system is listening and when it is not listening and convey relevance information to the users. The prototype also provides an interface that is understandable by digital photo users and enhances the structure of the photo browsing retrieval strategies to make them suitable for linking to a multimodal user interface. We have showed that the FlexPhoReS has yielded preliminary evidence revealing that FlexPhoReS already provides a good basis for supporting flexibility for 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 a 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 a more natural mode of interaction. At present, voice interaction is limited to a minimal vocabulary. This can be further expanded and developed to give the user a richer interaction experience.

One of the main thrusts for future work is implementing web based semi-automatic personal digital photo metadata annotation with multimodal interaction and integrating them into one system.

ACKNOWLEDGEMENT

The Malaysian Ministry of Science, Technology and Innovation (MOSTI) and Universiti Teknologi Malaysia (http://www.utm.my) kindly provided funding for the research through the E-Science MOSTI grant: 01-01-06-SF0244 (UTM-RMC Vot : 79226). A special vote of thanks is due to fellow research students and staff in the Research School of Informatics, at Loughborough University, Research Assistants Siti Azura and Shahrir and staffs in the Department of Computer Graphics and Multimedia.

REFERENCES

[1] N. A. Ismail, and A. O'Brien. 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., 2004.

[2] K. Rodden. How do people organise their photographs? BCS-IRSG Colloquium on Information Retrieval. <http://www.rodden.org/.../irsg.pdf>. [accessed 04.10.07] , 1999.

[3] K. Rodden and K. R. Wood. How do people manage their digital photographs? Proceedings of the SIGCHI conference on Human factors in computing systems. Florida. ACM, pp. 409-416, 2003.

[4] N. Van House, et al. From 'what?' to 'why?': the social uses of personal photos, <http://www.sims.berkeley.edu/.../vanhouse_et_al_2004a.pdf>, [accessed 06.10.07] ., 2004.

[5] Apple iPhoto. <http://www.apple.com/ilife/iphoto/>, [accessed 05.06.08] , 2008.

[6] Ulead. Ulead iMira. <http://www.imira.com/>, [accessed 04.06.06], 2004.

[7] Adobe Photoshop Album, [n.d.]. <http://www.adobe.com/products/photoshopalbum/main.html>, [accessed 04.06.06].Year

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

[9] A. Kuchinsky, et al. 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, 1999

[10] T.J. Mills, et al.. 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, 2000.

[11] J.C. Platt, et al.,. 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, 2003.

[12] H. Kang, and B. Shneiderman. 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, 2000.

[13] Flickr. <http://www.flickr.com/>, [accessed 12.06.08] , 2008.

[14] Picasa<http://picasa.google.com/>, [accessed 10.06.08] , 2008..

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

[16] S. Oviatt, et al. 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, 2000.

[17] R. Roope. Multimodal human-computer interaction: a constructive and empirical study. PhD thesis, Department of Computer Sciences., Tampere University, 1999.

[18] S. Oviatt. Multimodal interfaces. In: Jacko and Sears (eds.), The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications: Lawrence Erlbaum Associates, Inc, pp. 286-304, 2003.

[19] C. Benoît, and B. L. Goff. Audio-visual speech synthesis from French text: eight years of models, designs and evaluation at the ICP. Speech Communication, 26(1-2), 117-129 , 1998.

[20] S. Zhai, et al. Manual and gaze input cascaded (MAGIC) pointing. Proceedings of the Conference on Human Factors in Computing Systems (CHI'99). Pittsburgh. ACM, pp. 246-253 , 1999.

[21] T. Käster, et al. Combining speech and haptics for intuitive and efficient navigation through image databases. Proceedings of the 5th international conference on Multimodal interfaces. Vancouver. ACM, pp. 180-187, 2003.

[22] N.A. Ismail, and A. O'Brien. Multimodal Interaction for Visual Example Searching in Web-Based Personal Digital Photo Repository. 5th International Conference on Information Technology and Applications (ICITA 2008). Cairns, Queensland, Australia, 2008.

[23] N.A. IsmailFlexible Photo Retrieval (FlexPhoReS): a prototype for multimodal personal digital photo retrieval. PhD thesis, Research School of Informatics, Loughborough University , 2007.

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

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

[26] A. Hunt, 2004. Speech Recognition Grammar Specification Version 1.0, <http://www.w3.org/TR/speech-grammar/>, [accessed 15.05.08].

[27] SALT Forum, 2002. Speech Application Language Tags (SALT), <http://www.saltforum.org>, [accessed 10.04.06].