Speaking to See: A Feasibility Study of Voice-Assisted Visual Search

Victor Kaptelinin, Herje Wåhlen

Speaking to see: A feasibility study of voice-assisted visual search

Victor Kaptelinin[1,2], Herje Wåhlen[2]
[1] University of Bergen, Department of Information Science and Media Studies,
PO Box 7802, 5020 Bergen, Norway
[2] Umeå University, Department of Informatics, 901 87 Umeå, Sweden
vka062@uib.no, herjew@gmail.com

Abstract. The paper presents the concept, implementation, and a feasibility study of a user interface technique, named VAVS ("voice-assisted visual search"). VAVS employs user's voice input for assisting the user in searching for objects of interest in complex displays. User voice input is compared with attributes of visually presented objects and, if there is a match, the matching object is highlighted to help the user visually locate the object. The paper discusses differences between, on the one hand, VAVS and, on the other hand, voice commands and multimodal input techniques. An interactive prototype implementing the VAVS concept and employing a standard voice recognition program is described. The paper reports an empirical study, in which an object location task was carried out with and without VAVS. It was found that the VAVS condition was associated with higher performance and use satisfaction. The paper concludes with a discussion of directions for future work.

Keywords: Voice recognition, visual search, multimodal input, voice command.

1 Introduction

Visual search is a crucial component of a wide range of interactions between people and digital technologies; it involves scanning displayed information to detect the presence of an object of interest, indentify its location, or explore object's properties. For instance, if the user wants to make sure that the last email message from a certain customer has been actually answered, the user may scan the list of messages in the Inbox window and search for the last message from the client, visually locate the message line, and check whether the icon on the left contains a small arrow. Finding a certain street on a digital map, looking up information about a flight on a "Departures" monitor, and many other everyday interactions with electronic displays are critically dependent on visual search. Visual search may or may not involve carrying out an action with the object of interest.

In this paper we argue that for users of digital technologies visual search may be associated with certain problems, and that there is a need to provide the users with more advanced technological support for visual search. We introduce a user interface

technique, named VAVS (voice-assisted visual search), which aims to facilitate visual search by employing user's voice input for visually highlighting objects of interest.

In the remainder of this paper we present the rationale behind the VAVS technique, discuss how the technique is related to previous work, describe an interactive prototype of a system implementing the technique, and report a feasibility study, in which the prototype was employed in an object location task.

2 Background

Making a large number of information objects simultaneously available to the user for viewing has important advantages. In particular, it decreases the need for the user to open and "look inside" opaque containers, such as folders or pull-down menus to find objects of interest [2]. However, these advantages come with a price. In case of dense, complex displays, when the object of interest (the "target") is presented simultaneously with a large number of other objects ("distractors"), visual search becomes a more demanding task [9]. Problems with visual search can be aggravated by several factors, such as users' age (children and the elderly have more difficulties than young adults), level of stress, and certain health conditions, as well as how specifically the target is defined when a person carries out a visual search task (e.g., [4, 9]). The problems are likely to worsen in the future, since the screen size and resolution of computer monitors, public information displays, tabletops, and so forth, are ever-increasing, which means displaying more (and more complex) information objects.

Helping users visually identify their objects of interest has always been high on the agenda of the design of graphical user interfaces. Well-designed interfaces visually emphasize potentially important objects and de-emphasize less important ones [2, 9]. Relative visual salience of displayed objects can be a static feature of an interface or it can dynamically change depending on the task context (for instance, the default button in a dialogue window is highlighted to make it easier for the user to choose the most likely option).

These strategies for supporting users' visual search seem to have been often successful in the past and they remain to be useful. However, they are, arguably, not sufficient for addressing current challenges. Making potentially relevant objects visually salient does not scale up to complex displays and complex tasks. If all potentially relevant objects are visually salient, their absolute number can be overwhelming. In addition, when a large amount of information is displayed, it might be difficult for the system to anticipate just what objects can be of importance to a particular user in a particular context and, therefore, should be visually emphasized.

These logical arguments are consistent with the evidence obtained in empirical studies. For instance, Andrews et al. [1] describe "losing the cursor" and users' confusion caused by "windows and dialog boxes opening or gaining focus in unexpected locations" as common problems with large displays.

To address the problems, discussed above, we have developed a user interface technique for assisting the user in searching for objects of interest in complex

displays. The underlying idea of the technique, named VAVS (voice-assisted visual search), is employing user's voice input for guiding user's visual attention.

Figure 1 shows an overall structure of a VAVS-enabled interface. The user scans an image displayed on a screen (S) to locate a certain object. The user can also use a microphone (M) to describe object's attributes, such as its name. The voice input is processed and compared with attributes of objects displayed on the screen and, if there is a match, the matching object is visually highlighted. For instance, if a person, looking at a map of Colorado on a computer display, is saying "Hmm... Mancos… Mancos…", the location of the town on the map is temporarily highlighted.

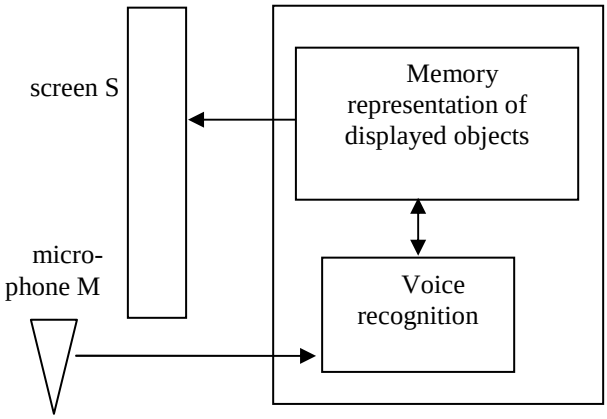

Fig. 1. Overall structure of a VAVS-enabled interface.

The VAVS technique should be differentiated from two other ways of using user's voice input, which have been actively explored in previous research: voice commands (in a broad sense, including voice-based queries) and multi-modal input techniques.

Like voice commands, which are an increasingly common interaction technique, for instance, in in-car systems [5], VAVS also employs users' voice input. Unlike voice commands, however, VAVS does not cause substantial changes in the state of the system. Its effect is limited to visually highlighting potential objects of interest. If a user's voice input results in highlighting some other object than the desired one (either because of a user's mistake or system's misinterpretation) the user can simply ignore the highlighting when proceeding with their task. It also means that VAVS users do not have to be overly concerned about negative consequences of their mistakes (while users of voice command systems have to overcome a substantial initial barrier before they start to feel comfortable with a system [5]).

A related approach to employing user's voice in human-computer interaction is supporting multi-modal input, or multi-modal dialogue, that is, enabling the use of voice input in combination with other interaction modalities [3,7,8]. An example of this approach is the classic "Put-that-there" system [3], which combines voice and gesture. For instance, to move an object across a large display the user specifies a command by voice (i.e. by saying "put"), points to an object and selects it by saying "that", and finally indicates a new location by pointing to it and saying "there".

Users of the “Put-that-there” system, as well as users of more recent systems that implement the same general approach [8], need to know—in advance—the spatial locations of objects of interest and convey these spatial locations to the system when instructing it to carry out a desired action. Support for selecting an object to indicate the system what it should act upon (cf. Windows Speech Recognition [10]) may partially overlap with support of visual search, but the general approach adopted by VAVS is, in a sense, opposite. According to that approach, it is the system that conveys the spatial locations of objects of interest to the user, rather than the other way around. Accordingly, a VAVS system has a number of features differentiating it from multimodal input systems. For instance, all potential objects of interest, rather than just potential objects of actions, should be “highlightable”.

The next section presents a feasibility study intended to gain empirical evidence on whether VAVS can be helpful when supporting users in finding objects of interest on complex displays.

3. Method

Participants. Eight university students, native Swedish speakers and fluent English speakers, 23 to 33 years old, took part in the study.

Procedure. The participants were tested individually. Each session started with a profile calibration procedure that took five to twelve minutes. After that each participant was presented with a series of object location tasks. In each task a participant was presented with a name of a map region in the top left corner of the screen and was required to locate and click the corresponding map region using the mouse. The user had to click the correct map region to proceed to the next task. Each participant was presented with 96 object location tasks divided into two blocks. One of the blocks corresponded to the “VAVS” condition (voice input was enabled), and the other block corresponded to the “non-VAVS” condition (voice input was disabled). In each block the first five tasks were practice tasks, not included in the analysis. Finally, the participants were briefly interviewed about their experience with VAVS. The duration of a typical session with a participant was about 30 min.

Equipment. The hardware used in the study was an Apple MacBook Pro computer (15-inch, 2.33 GHz Intel Core 2 Duo processor, 4 GB SDRAM) running Mac OS X 10.6.3, connected to two external devices: Microsoft IntelliMouse Explorer 3.0 and Logitech USB Desktop Microphone.

Prototype. An interactive prototype of a VAVS-enabled system was developed for the study in AppleScript and JavaScript. The prototype was integrated with a speech recognition program, Nuance MacSpeech Dictate International, version 1.5.8. The visual interface was implemented as an HTML document opened in full screen mode.

The functionality of the prototype included: (a) displaying a map featuring a number of regions (“countries” or “states”), (b) displaying the name of one of the map regions in the top left corner of the screen, (c) measuring the time interval between presenting a name of a region and a mouse click on the corresponding map region, (d) recognizing a map region name uttered by the user, and (e) visually highlighting the

map region corresponding to the name. In the control ("non-VAVS") condition functions (d) and (e) above were disabled.

Materials. Two maps, loosely based on Adobe Photoshop filter-generated images as reference for map region borders, were created for the study. *Map A* was derived from a map of Europe, and real English names of European countries were randomly assigned to different map regions (see Figure 2). *Map B* was derived, in a similar manner, from a US map.

The maps were designed to make sure the participants were familiar with the names of the map regions but could not use their previous knowledge to infer the locations of map regions from their names. Therefore, the participants had to visually scan the maps in order to complete the experimental tasks.

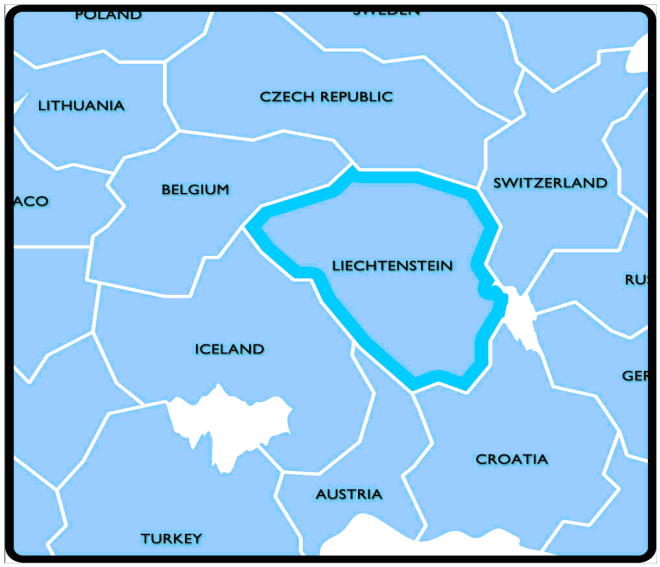

Fig. 2. An adapted fragment of Map A ("Liechtenstein" is visually highlighted).

Design. The study employed a one-factor within-subject design, with the independent variable being Voice Input ("VAVS" condition vs. "non-VAVS" condition). The main dependent variable was task completion time.

The design was balanced to minimize the potential effects of condition sequence and map types. The participants were divided into two equal sub-groups. The first sub-group completed the first block of tasks in the "VAVS" condition and the second block in the "non-VAVS" condition; for the second sub-group the sequence was the opposite. In each of these two sub-groups half of the participants worked with Map A in the "VAVS" condition and Map B in the "non-VAVS" condition, while for the other half the correspondence between maps and conditions was the opposite.

4. Results

As mentioned, the experiment procedure required a task to be correctly completed before the next task could be presented. All participants were able to complete all tasks in both conditions, which allowed us to use time to correctly complete a task as an integral performance indicator, in which error costs, both participants' mistake and voice recognition errors, were reflected as added "error time".

Voice recognition error rate in the VAVS condition—calculated as the percentage of tasks, in which the participants had to pronounce a state or country name more than once—was 19%. In two cases the experimenter had to intervene and suggest the right pronunciation (while the tasks were performed by the participants themselves). A likely reason for the high error rate was that native Swedish speakers were asked to pronounce English words.

Figure 3 shows accumulated times for completing blocks of tasks in the two experimental conditions for each of the eight participants. Figure 3a shows the results of the four participants (S1, S2, S3, and S4), who worked with Map A in the "VAVS" condition and Map B in the "non-VAVS" condition. Figure 3b shows the results of the four participants (S5, S6, S7, and S8), who worked with Map B in the "VAVS" condition and Map A in the "non-VAVS" condition.

The results, shown in Figure 3, indicate that in the "VAVS" condition *each* of the participants completed the experimental tasks faster than in the "non-VAVS" condition. While the average accumulated task completion time in the "non-VAVS" condition was **384** seconds; in the "VAVS" condition it was **176** seconds.

The results were analyzed using the Wilcoxon signed-rank test. The difference between the "VAVS" and "non-VAVS" condition was found to be statistically significant ($N=8$, $W_+=36$, $W_-=0$, $p=.005$).

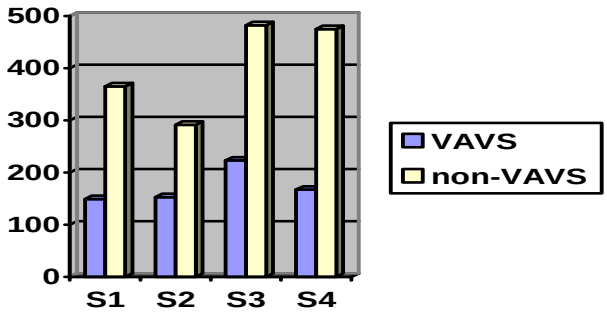

Fig. 3a. Accumulated task completion times, in seconds, for the experimental conditions of the study. Participants: S1, S2, S3, and S4. "VAVS": Map A, "non-VAVS": Map B.

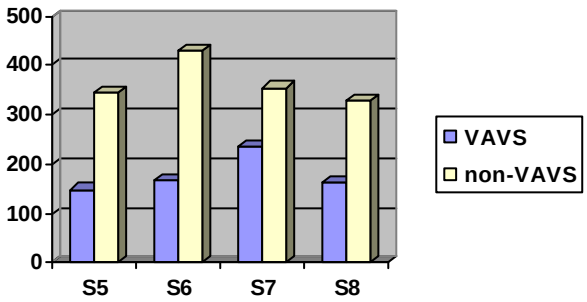

Fig. 3b. Accumulated task completion times, in seconds, for the experimental conditions of the study. Participants: S5, S6, S7, and S8. "VAVS": Map B, "non-VAVS": Map A.

In their interview comments all participants indicated that they were positive about the VAVS technique and wanted it to be used in a diversity of everyday contexts.

5. Discussion of results and future work directions

The results of our study suggest that employing user voice input for visually highlighting objects of interest can be associated with higher performance and positive user experience. Given that the use of voice at the user interface is complicated by a number of factors [2], and speech-based interfaces have been, in general, much less successful than it was anticipated in the past [6,7], we consider the results of our study encouraging. The study also showed that a standard speech recognition program can be accurate and reliable enough to support VAVS-enabled interaction.

It should be noted that advantages of VAVS were observed in conditions in which the advantages were not self-evident. The participants had to speak a foreign language, which was probably one of the reasons behind the high voice recognition error rate and, consequently, resulted in higher task completion times in the VAVS condition. In addition, the map image used in the study was relatively simple, which meant that unassisted visual search remained a viable option. It is reasonable to assume that if the users spoke their native language and worked with large displays and complex images, VAVS' advantages would be even more significant.

Can the findings be explained by a "negative familiarity" effect, that is, by target familiarity being an impediment rather than help in the specific task used in the study? If this explanation is correct, the findings from our study are only valid for rare instances of search tasks. However, the results do not support this hypothesis: if it were correct, the longest search time would be for "Sweden", which was participants' home country. In fact, the average search time for "Sweden" was shorter than for any other country name used in the experiment.

The study reported in this paper is a feasibility study, an initial phase of exploring the VAVS technique. Choosing unassisted visual search as a baseline for comparison

was a natural choice for this first step. Further exploration of the technique is planned to compare VAVS with other types of visual search support, such as using text search strings for visually locating objects displayed on the screen. Other possible issues to be explored in future research are as follows:

Augmented reality applications. In augmented reality applications VAVS can be used to help people locate objects of interest in the physical environment. For instance, providing voice input to a wearable system that includes a head up display can help a supermarket customer locate a certain product on a shelf.

Using small screen devices to view large images. Visual search can be especially difficult if the user scans a large image (e.g., a map) using a small screen device, such as a smartphone. A variation of VAVS can be implemented to recognize user voice input and, if it matches an object, which is a part of the large image but not displayed in the small window, indicate the direction in which the window needs to be scrolled to display the object.

2D sound feedback. A potential problem with VAVS is that in case of very large, complex, and dynamic displays the visual highlighting produced by VAVS could be difficult to detect. A possible solution to this problem is to supplement visual highlighting with a 2D sound signal that would direct user's attention to the general spatial location of the object of interest.

References

1. Andrews, C., Endert, A., and North, C. Space to think: Large, high-resolution displays for sensemaking. In Proc. CHI 2010. ACM Press, 55-64 (2010).
2. Benyon, D., Turner, P., and Turner, S. Designing Interactive Systems: People, Activities, Contexts, Technologies. Addison-Wesley, NY (2005).
3. Bolt, R. A. "Put-that-there": Voice and gesture at the graphics interface. In Proc. of the 7th annual conference on Computer graphics and interactive techniques. ACM Press, 262-270 (1980).
4. Fabiani, M., Low, K. A., Wee, E., Sable J. J., and Gratton, G. Reduced Suppression or Labile Memory? Mechanisms of Inefficient Filtering of Irrelevant Information in Older Adults. J. Cogn. Neurosci., 18: 4, 637-650 (2006).
5. Lau, T. and Reed, D. Speech-activated user interfaces and climbing Mt. Exascale. Communications of the ACM, 52: 6, 10-11 (2009).
6. Manaris, B. Natural Language Processing: A Human-Computer Interaction Perspective. Advances in Computers, 47, 2-68 (1998).
7. Nielsen, J. Voice Interfaces: Assessing the Potential, http://www.useit.com/alertbox/20030127.html
8. Voida, S., Podlaseck, M., Kjeldsen, R., and Pinhanez, C. A study on the manipulation of 2D objects in a projector/camera-based augmented reality environment. In Proc. CHI 2005. ACM Press 611-620 (2005).
9. Ware. C. Information Visualization. Perception for Design. Second edition. Morgan Kaufmann, Amsterdam (2004).
10. What can I do with Windows Speech recognition?, http://windows.microsoft.com/en-US/windows7/What-can-I-do-with-Speech-Recognition