Multi-modal Interactions for Inconspicuous Mobile Eye-Tracking

Niaja Farve
Fluid Interfaces Group
MIT Media Laboratory
20 Ames Street
Cambridge, MA
nfarve@mit.edu

Pattie Maes
Fluid Interfaces Group
MIT Media Laboratory
20 Ames Street
Cambridge, MA
pattie@media.mit.edu

Copyright is held by the owner/author(s).

CHI 2014 Workshop on Inconspicuous Interaction, April 26, 2014, Toronto, ON, Canada.

Abstract
Multi-modal interactions have long been explored as a way to improve user experience. With the development of mobile eye-tracking technology, the advantages of using multiple input modes can be applied to mobile eye-tracking applications. The paper introduces a platform for development of mobile eye tracking called Ogle; along with a key feature it possesses the ability to develop multimodal inconspicuous applications. Similarly this paper explores the advantages of using multi-modal interactions with mobile eye-tracking technology through development of two early-stage applications. By using both gaze and audio as input the complexity of the inconspicuous system could be increased without sacrificing user experience.

Author Keywords
Human Computer Interaction; Mobile Eye-Tracking; Multi-Modal Interactions

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction
Although the eye, its movements, and functionality have been studied since the late 1800's [1], studies involving eye tracking have been primarily constrained

Figure 1. Pupil eye-tracking headset with two cameras

to stationary applications, due to limitations in equipment. Recent advances in hardware technology have made it possible to use eye-tracking technology in a mobile context. This new application of eye-tracking technology brings new challenges, as users need to be able to give and receive real-time information quickly without disrupting their current activity. This paper explores how to offset the limitations of gaze input; auditory input can be used to form multi-modal human computer interactions.

Our contribution is to introduce a work-in-progress platform called Ogle easy to use platform for the development of mobile inconspicuous eye-tracking applications that allows for quick, versatile, and effective applications for mobile eye-tracking applications. A unique feature of this platform is the ability to use multi-modal interactions in combination with eye-tracking technology. Users can define modes of input such as gaze, audio, or sensors. These input modes can then be tied to triggers to process information received from any single or combination of modes. Through the use of gaze and audio input a wider array of use cases are supported without decreasing ease of use, a relationship that is usually inversely related. By introducing multi-modal interactions to mobile eye-tracking, applications that would normally be impossible with only one mode of input become easy and intuitive. Several advantages become apparent through the use of multi-modal interactions.

Heads Up/Hands Free Interactions

One of the strongest advantages of using multiple input methods is the ability to make human-computer interaction through eye-tracking hands free and heads up. While several devices have been made to incorporate gaze, the user is limited in what information he or she can request and what information is presented. To give the user more means of expression, typically secondary devices, such as cell phones, are introduced. This causes the user to switch visual attention from the area of interest to the secondary device. This becomes time consuming and makes input from gaze less effective and unnecessary.

Two Modes of Input

The second advantage of using multi-modal interactions is the ability to interchange input methods. Gaze and auditory input methods both provide limitations. Gaze only provides visual information. To interpret or use this visual information in any way other methods of input are required. Auditory input allows the user to give specific commands, but these commands are limited to information the user or the database already posses. Auditory feedback also allows the user to receive information while remaining heads-up and hands free. By using both gaze and audio as modes of input, the user can overcome the shortcomings of one mode of input by complementing it with the other. Once this visual information is gathered the user can control how it is queried and what information is given back through voice commands. The user is ultimately able to perform more actions seamlessly using interactions that are intuitive.

Easily Expandable

Finally, a third advantage is that the use cases for multi-modal interactions are easily added upon by simply adding to the database that is queried upon. To add more features the database of commands simply needs to be expanded. If only one mode of interaction

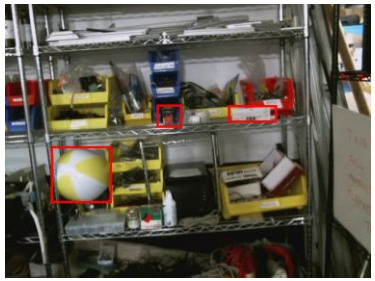

Figure 2. Field of view with three items from database detected.

were available, complex ways of interaction would need to be developed in order to access these features. For example, if gaze were the only input method, a query on in an image would need to be activated through the use of eye gestures. For each new action a new gesture must be created. This cohesive use of both modes of input adds additional features without substantially adding complexity to the system or its use. Rather than having to learn new gestures or use an external device the user simply needs to learn the name of new command added to the system.

Previous Work

In recent years, multimodality has been deftly explored in mobile phone applications. Balmund's research both voice and graphics were used to support multimodal input and show an improvement in user experience [2]. Baillie's research in user interactions using multimodal applications found that multimodality significantly enhanced the applications and users found it particularly effective when participating in mobile contexts (e.g. while walking) [3]. Both of these works show that multimodality improves applications and user experiences. These advantages are even more present when multimodal interaction is used for mobile applications. From these works we postulate that utilizing multimodality for mobile eye-tracking applications will not only improve applications used for such a system but also drastically improve applications' effectiveness.

Wang's work on integrating with eye-gaze, voice, and manual response showed that the use of gaze and voice was natural and intuitive than manual input [4]. Both of the works done by Balmund and Hatfield have developed models, principles, and guidelines to integrate voice and eye-tracking technology [5]. However, most of these models have only been applied to stationary applications.

Current work

To test the effectiveness and ease of use of Ogle, a handful of simple applications have been developed using the Pupil headset. Pupil is an eye-tracking hardware and software platform [6] that uses an open source headset and software. With Pupil a user employs two cameras (a "world" camera and an "eye" camera) that communicate and produce videos displaying the users field of view and gaze direction. Using this system two separate use case have been explored.

The first application involves using the gaze direction and the field of view image to capture snapshots. A snapshot can be taken of the area of interest when the gaze location stays relatively the same for more than a few seconds. This snapshot can be triggered and processed appropriately based on the voice command given by the user. Currently the Google speech to text API is used to execute voice commands. This method of using both voice and gaze position allows the user to gain more information about an item quickly. These types of applications could be especially useful for accessibility or hands free applications. Other possible commands include "remember" to capture an image and save it, "read" to read text aloud, and "search" to search an image or text through online search databases.

The second application involves known objects. In this case the field of view is constantly being searched for objects in the database. If one is detected the user can request more information about the object using the

voce command "more". This is particularly useful when a user is in a particular location like a grocery store or on a city tour in which case the objects the user will likely interact with are known ahead of time. Each object can be tied to information, such as nutritional or historical information. If multiple known objects exist in the field of view the user can distinguish between them using their gaze location. This use case has been demonstrated using three objects; a beach ball, a panic button, and a circuit board. A cascade classifier was made for each object using OpenCV [7]. When either of these items exists in the field of view the user is given an auditory cue. The user can then chose to request more information about an object. If the user gives the appropriate command the user is read the information about that object. A user can continue to give the command while the objects are in the field of view. The total time from command recognition to auditory feedback averages 3 seconds; again we believe this can be sped up significantly with minor improvements.

Future Work

In order to make Ogle useable and more effective, more improvements and expansions can be made.

Completely Mobile System

While this system of multi-modal input is effective for both sedentary and mobile applications it is particularly useful in a mobile scenario. In such a scenario it is more useful to not have to depend on an additional device or have the user take their attention away from an object of interest. Future work will be done to incorporate this system on a pocket PC such as Intel's Galileo or a mobile device.

Increasing Database Size

In order to test the speed and robustness of the system a larger database of images will need to be used. A larger database will give the user more options but may introduce a delay as each image is searched. Utilizing existing specialized databases can help test the effectiveness in pre-determined locations, such as supermarkets.

More Commands

Finally, work will be done to increase the list of commands available to the user. Adding more commands will help test if complexity can be added to the system without drastically increasing complexity of use. User input will be used to determine what information is most pertinent therefore defining what commands will be added. The voice to text system will also be converted to a native system so that the Internet is not necessary to recognize and process commands.

Conclusion

Developments in wearable technology have made it possible to use eye tracking for mobile applications. However, mobile eye tracking has its limitations. Using gaze as a sole input method limits a system's use. Traditionally to compensate for these limitations secondary devices have been added to the system. This however, forces the user to change focus of attention. By using a multi-modal voice-gaze approach, this paper demonstrates that a user can access relevant information quickly and easily without changing gaze locations.

Acknowledgements

Thanks to Moritz Kassner and Patera William for developing the PUPIL system and for making it open source.

References

[1] Roper-Hall, Gill (2007), Historical Vignette Louis Emile Javal (1839-1907): The Father of Orthopics, in American Orthoptic Journal, January 1, 2007, 131-136.

[2] Schatz, Raimund (2005), HCI Proceedings Volume 2, Developing Mobile Multimodal Applications.

[3] Baillie, L., Schatz, R., Simon, R., Anegg, H., Wegscheider, F., Niklfeld, G., Gassner, A.: Designing Mona: User Interactions with Multimodal Mobile Applications. In: Proceedings of 11th International Conference on Human-Computer Interaction (HCI International), pp. 22–27. Lawrence Erlbaum Associates (2005)

[4] J. Wang, "Integration of eye-gaze, voice and manual response in multimodal user interface," in Proc. IEEE Int. Conf. Systems, Man, and Cybernetics. 1995, pp. 3938-3942

[5] F. Hatfield, E. A. Jenkins, M. W. Jennings, and G. Calhoun. "Principles and guidelines for the design of eye/voice interaction dialogs," in Proc. The 3rd Symposium on Human Interaction with Complex Systems (HICS '96), 1996, pp. 10-19.

[6] Kassner, Moritz. (2012) William, Patera, PUPIL Constructing the Space of Visual Attention. (Master's thesis).

[7] "Cascade Classifier Training." Cascade Classifier Training — OpenCV 2.4.8.0 Documentation. N.p., n.d. Web. 07 Jan. 2014.