eyeLook: Using Attention to Facilitate Mobile Media Consumption

Connor Dickie, Roel Vertegaal, Changuk Sohn and Daniel Cheng

Human Media Lab, Queen's University
Kingston, ON. K7L 3N6, Canada
{ connor, roel, sohn, cheng } @ cs.queensu.ca

ABSTRACT

One of the problems with mobile media devices is that they may distract users during critical everyday tasks, such as navigating the streets of a busy city. We addressed this issue in the design of eyeLook: a platform for attention sensitive mobile computing. eyeLook appliances use embedded low cost eyeCONTACT sensors (ECS) to detect when the user looks at the display. We discuss two eyeLook applications, seeTV and seeTXT, that facilitate courteous media consumption in mobile contexts by using the ECS to respond to user attention. seeTV is an attentive mobile video player that automatically pauses content when the user is not looking. seeTXT is an attentive speed reading application that flashes words on the display, advancing text only when the user is looking. By making mobile media devices sensitive to actual user attention, eyeLook allows applications to gracefully transition users between consuming media, and managing life.

ACM Classification Keywords

H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces: Input devices and strategies.

Keywords

Attentive User Interfaces, Ubiquitous Computing, Context-Aware Computing, Eye Tracking, Mobile Computing.

General Terms

Design, Human Factors.

INTRODUCTION

With the introduction of ubiquitous mobile computing, human-computer interaction has fundamentally changed from a stationary one-to-one, to a mobile one-to-many relationship. This means the consumption of computer media is shifting away from relatively controlled environments such as the office or home, to more unpredictable environments, such as public transit and the outdoors. Despite this shift in usage circumstance, dialogue with mobile devices remains based on the notion that a user's primary task is to access computer resources and, that these computer resources may monopolize user attention. As a consequence, user attention has become a limited resource, which is continually vied for by various devices and life activities. It is our belief that the proliferation of ubiquitous mobile digital devices necessitates a new way of thinking about human-computer interaction. Unlike traditional tools, computers are active communicators. However, they tend not to negotiate their communications in a manner consistent with human social norms. Consider how often people are interrupted by a mobile phone because it rings without any regard for their current activity. This example illustrates a serious underlying flaw in user interfaces; the computer's lack of knowledge about the task focus of its user. These failings are exacerbated when the user is in a mobile context, where the primary task focus may be the navigation of a street, or taking part in a social activity [4]. We addressed this issue in the design of eyeLook: an attentive mobile media device that senses user attention in order to optimize primary task focus. We discuss two eyeLook applications, seeTV and seeTXT, that facilitate courteous media consumption in mobile contexts by dynamically responding to the available attentional resources of a user. seeTV is a mobile attentive video player that automatically pauses content when the user is not looking. seeTXT is a mobile attentive speed reading application that flashes individual words on the display, advancing text only when the user is looking. This behavior allows eyeLook to explicitly negotiate the timing of communications with the user, pending their current needs. In eyeLook we modeled our design strategy on the most striking metaphor available: that of human group communication [10].

UIST'05, October 23-27, 2005, Seattle, Washington, USA.

Attention and Turn Taking

In human conversation, attention is inherently a limited resource. Humans can only listen to, and absorb the message of one person at a time. In meetings, listener attention is optimized by allowing only one person to speak at a time. This remarkably efficient process of turn taking uses nonverbal cues such as eye contact to convey attention between interlocutors [9]. In group conversations, eye contact indicates with about 80% percent accuracy whether a person is being spoken or listened to in four-person conversations [9]. When a speaker falls silent, and looks at a listener, this is perceived as an invitation to take the floor. Vertegaal [9] showed that in triadic mediated conversations, the number of turns drops by 25% if eye

contact is not conveyed. Humans use eye contact to optimize the turn taking process for three reasons:

1. Eye fixations provide the most reliable indication of the target of a person's attention, including their conversational attention [9].
2. The perception of eye contact increases arousal, which aids in proper allocation of brain resources, and in regulating inter-personal relationships [9].
3. Eye contact is a nonverbal visual signal, one that can be used to negotiate turns without interrupting the verbal auditory channel.

Turn taking provides a powerful metaphor for the regulation of communication with ubiquitous mobile devices. By incorporating eye contact sensing into mobile devices, we give them the ability to recognize and act upon innate human nonverbal turn taking cues.

Eye Contact Sensing

Our eyeCONTACT sensor (ECS) [8] consists of a wireless camera that sends video using a 2.4Ghz transmitter, to a PC running ECS server software that finds pupils within its field of view using computer vision (Figure 1). A set of infrared LEDs is mounted around a camera lens. When flashed, LEDs produce a bright pupil reflection (red eye effect) in eyes within range. Another set of LEDs is mounted off-axis. Flashing these produces a similar image, with black pupils. By syncing the LEDs with the camera clock, a bright and dark pupil effect is produced in alternate fields of each video frame. Video frames are wirelessly transmitted to a PC where a simple algorithm finds any eyes in the frame. This is accomplished by subtracting the even and odd fields of each frame, leaving only the pupils remaining. The LEDs also produce a reflection on the surface of the eyes. These glints appear near the center of the detected pupils when the onlooker is looking at the camera, allowing the detection of eye contact without any calibration. ECS stream information about the number and location of pupils, and whether these pupils are looking at the device over a TCP/IP connection. This specific prototype can reliably sense eye contact at up to 1 meter.

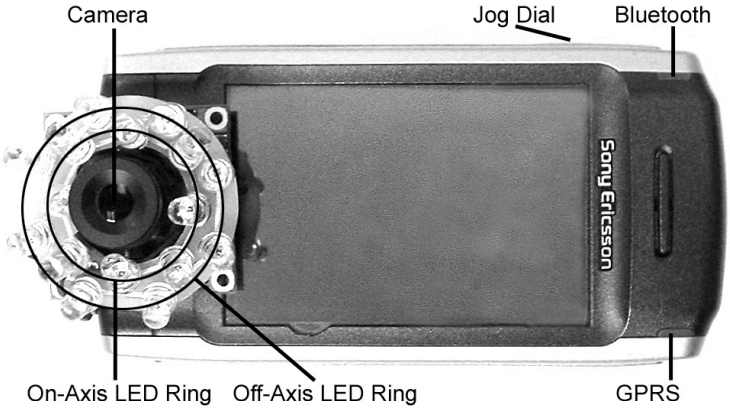

Figure 1. eyeLook with integrated ECS

PREVIOUS WORK

Attentive interfaces cannot exist without knowing the attentive state of a user. LAFCam [2] makes use of the involuntary attentive cues videographers utter. It uses an AI model to recognize voice and laughter. LAFCam was used to successfully find interesting moments in a video based on the nonverbal utterances made by the videographer during filming. Auramirror [5] is a video mirror that renders the virtual attentive auras that encompass groups of people during conversations by superimposing animated bubbles of attention over participants' heads. Auras grow towards interlocutors to form tunnels of attention. When interlocutors look at Auramirror to see the tunnel, it collapses as the target of their visual attention changes. With the GAZE-2 [10] project, Vertegaal demonstrated that eye tracking is a reliable channel of input for determining the attentive state of a user. Using eye trackers, GAZE-2 observes who participants look at during mediated group conversations. By automatically rotating 2D video images of individuals toward the person they look at, participants in a 3D meeting room can see who is talking to whom. Eye-aRe is an early attention sensor [6] in the form of augmented glasses that detect when the wearer is looking in the direction of another device or another user augmented with Eye-aRe technology. Eye-aRe detects the light emitted from other Eye-aRe sensors as well as pauses in user's eye movements. Similarly, the Attentive Cell Phone [8] used a low-cost ECS and speech analysis to detect whether its user was in a face-to-face conversation. This information is communicated to callers in an Instant Message interface to allow them to employ basic social rules of interruption.

EYELOOK IMPLEMENTATION

The eyeLook platform consists of a SonyEricsson P900 smartphone [7] running the Symbian 7.1 UIQ operating system. The smartphone has been augmented with a low-cost, wireless ECS that is situated to receive direct eye gaze when users look at the display. eyeLook allows applications to communicate wirelessly with the ECS server using TCP/IP over built-in Bluetooth or GPRS radios. The 5-way Jog Dial on the side of the phone allows for single-handed manual operation of applications. Media files can be stored internally, or on a MemoryStick Duo.

seeTV

seeTV is an application for eyeLook that plays video only when the user looks at the display. The seeTV application consists of three components; a telnet module that connects to the ECS over Bluetooth or GPRS, logic to determine if any eyes are looking in the direction of the device and Jog Dial events. An MPEG4 playback engine plays or pauses video on the command of the logic component. Manual

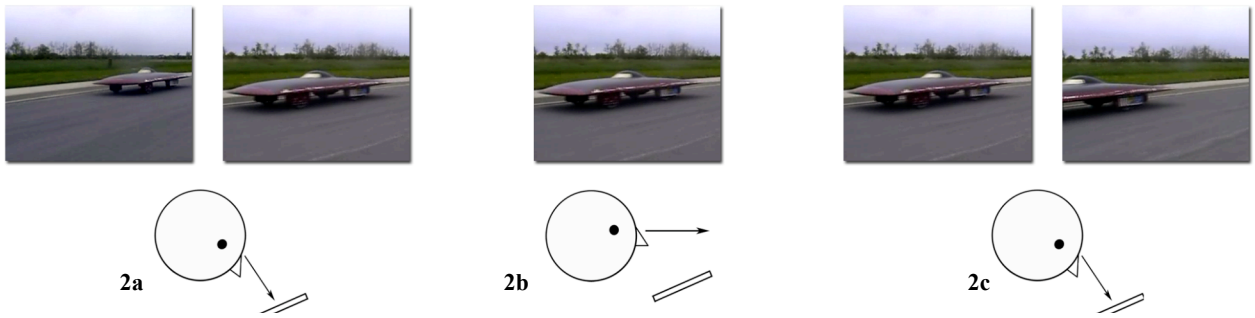

Figure 2. eyeTV operation

input from the Jog Dial produces different results depending on the context. Upwards or downwards 'dialing' modulates sound loudness while viewing content. 'Pushing' the Jog Dial away from the user while viewing content closes the Video View and opens the Video Index. Dialing moves the cursor up or down in the index. 'Clicking' the Jog Dial in opens the highlighted video in a new Video View. P900 video limitations require a trade-off between high frame-rate and high-resolution video. We prepare video content for eyeLook using Quicktime. Prepared files typically have a resolution of 176x144 pixels and refresh at 12 frames per second. Audio consists of a single 8kHz channel encoded from 16bit samples. Total data rate with these settings does not exceed 120 kbits per second, which allows a 2-hour movie to fit in about 85MB of storage.

seeTV Scenario

Like many, Jason spends a portion of his day commuting to work. The commute is mostly spent waiting, either for a train to arrive, or to travel to a particular station. There are, however, a few moments along the commute where Jason needs to be very alert. These moments usually occur when he is navigating his way to the next train or buying coffee, but can occur at other unpredictable times. Jason is already watching a video on his eyeLook (see Fig. 2a) when we find him standing in a busy line to buy coffee. The barista yells "Next!", and the video pauses while Jason orders, pays for, and receives his coffee in his free hand (see Fig. 2b). Because seeTV responded to his eye-gaze, the video resumes from the exact frame he last saw before he looked at the barista to order his coffee (see Fig. 2c). Jason finds a place to sit, and puts his eyeLook on his lap. He resumes watching video while holding his coffee in both hands to keep them warm. When the train arrives, he takes eyeLook in his free hand and enters the train. There is little space, so Jason must stand. He elects to put his eyeLook into his pocket so he can have a free hand in order to grasp an overhead handle. Many people debark the train at the next station and Jason is able to find an empty seat. He sits and promptly removes his eyeLook from his pocket and resumes where he left off, watching without interruption until the video ends. Still sipping his coffee, Jason presses back on the Jog Dial and closes the Video View revealing the Video Index. He dials the Jog Dial upwards to highlight a desired video file. He opens it in a new Video View by clicking the Jog Dial in, and begins watching.

seeTXT

seeTXT is a speed reading application that presents a body of text one word at a time in a stationary location on the display. Text is advanced only when a user looks at the display. The Rapid Serial Visual Presentation (RSVP) technique first studied by Forster [3] is an ideal candidate for overcoming the limitations of small displays [1]. RSVP text trades time for space. The effect of this is that text becomes dynamic as the display is constantly updated with the next word in the corpus. Eyes become static, as they are freed from the requirement to saccade across the page. Text can be larger and of higher resolution as only one word needs to fit on the display at a time. The seeTXT application has two components; a telnet module that connects to the ECS over Bluetooth or GPRS, and logic that enables seeTXT to determine how many eyes are looking in the direction of the device. The logic component also sends appropriate messages to the text display module. The text engine refreshes words on the screen only when a user is looking. As a security measure, displayed text will have a red tint if more than one pair of eyes is detected looking at the screen. While the user is looking, up or down dialing of the Jog Dial increases or decreases the rate at which words refresh on the screen. Clicking-in the Jog Dial manually halts advancing text. Looking at the display or clicking-in the Jog Dial continues text advancement. Pushing away on the Jog Dial when in Text View closes the Text View and opens the Text Index. Dialing up or down scrolls through the user's list of text files. Clicking in the Jog Dial opens a new Text View with the currently selected file as its contents.

seeTXT Scenario

As a Law Student in her final year of study, Fatima has what seems to be an inexhaustible supply of reading material. Before she leaves the library, Fatima emails a few text files to her eyeLook. A few moments later she feels a vibration in her pocket that notifies her that the email, and files, have been transferred to her eyeLook. It is a rather long, but unremarkable walk to the grocery store where Fatima is headed next. She initiates the seeTXT

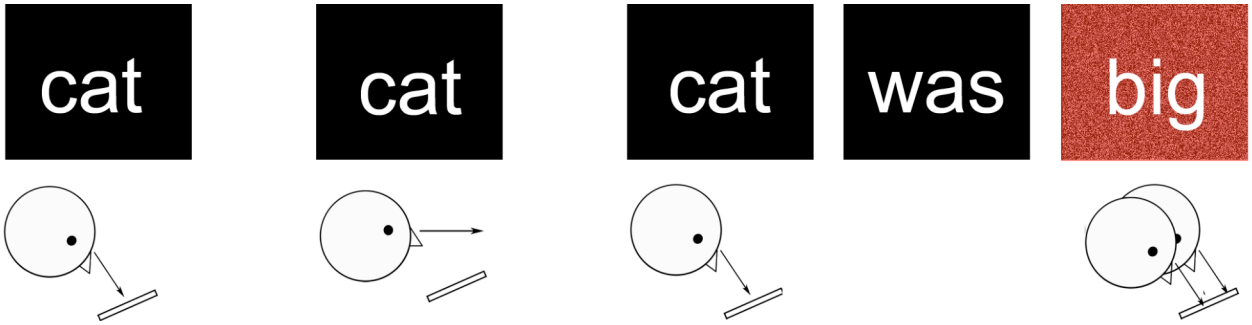

Figure 3. seeTXT operation

application and opens the first of a list of text files. Individual words from the text rapidly flash on the screen in a large high-resolution font (see Fig. 3a). Fatima, an adept eyeReader, uses the Jog Dial to increase the text refresh rate. While walking, Fatima holds the eyeLook away from her face so she can better survey her environment. Despite the small screen size, the text is large enough for her to easily read displayed words. The automatic pausing ability afforded by seeTXT allows her to frequently assess her proximity to any obstacles she may encounter while walking, without losing her place in the text (see Fig 3b). As Fatima approaches the grocery store, she slips her eyeLook into her bag. After shopping, Fatima decides to take public transit instead of walking home. After finding a seat on the bus, Fatima finally has an opportunity to continue with her reading. She reaches for her eyeLook and continues where she left off (see Fig. 3c). Suddenly the screen is washed in a red tint (see Fig. 3d). She quickly covers the screen knowing that someone else may be reading her confidential legal documents.

Future work

Having more than one eyeLook would allow users to have parallel access to more than one application. Wireless input devices such as keyboards and mice could be shared across each separate device, with them affecting only the device that is currently in user focus. As the user's focus switches from eyeLook to eyeLook, data from the input devices follows his focus. Handling punctuation and pauses in seeTXT would create a better user experience [1]. Incorporating a feedback-loop to automatically control the refresh-rate of advancing seeTXT text would free the user from having to manually dial any speed adjustments.

Conclusions

In this paper we presented eyeLook: a platform for attention sensitive mobile computing that uses an embedded low cost ECS to detect when the user looks at a mobile display. We discussed two applications, seeTV and seeTXT, that facilitate courteous media consumption in mobile contexts by using information from the ECS to dynamically respond to the user attention.

REFERENCES

1. Laarni, J. "Searching for Optimal Methods of Presenting Dynamic Text on different Types of Screens." In Extended Abstracts of NordiCHI 2002. Århus, Denmark, ACM Press 2002.
2. Lockerd, A. and F. Mueller. "LAFCam: Leveraging Affective Feedback Camcorder." In Extended Abstracts of CHI 02. Minneapolis: ACM Press, 2002, pp. 574-575
3. Mills, C., Weldon, L. "Reading Text From Computer Screens." ACM Computing Survey. ACM Press 1988.
4. Oulasvirta, A., Salovaara, A. "A Cognitive Meta-Analysis of Design Approaches to Interruptions in Intelligent Environments." Extended Abstracts, CHI 2004. Vienna, Austria.
5. Skaburskis, A. W., Shell, J.S., Vertegaal, R., and Dickie, C. "AuraMirror: Artistically Visualizing Attention." In Extended Abstracts of ACM CHI 2003 Conference on Human Factors in Computing Systems, 2003.
6. Selker, T. et al. "Eye-aRe, a Glasses-Mounted Eye Motion Detection Interface." In Extended Abstracts of CHI 2001. Seattle: ACM, 2001.
7. Sony Ericsson, http://www.sonyericsson.com
8. Vertegaal, R. Dickie, C., Sohn, C, and Flickner, M. "Designing Attentive Cell Phones Using Wearable EyeContact Sensors." In Extended Abstracts of ACM CHI 2002 Conference on Human Factors in Computing Systems. Minneapolis: ACM Press, 2002.
9. Vertegaal, R. and Ding, Y. "Explaining Effects of Eye Gaze on Mediated Group Conversations: Amount or Synchronization?" In Proceedings of CSCW 2002 Conference on Computer Supported Collaborative Work. New Orleans: ACM Press, 2002.
10. Vertegaal, R. Weevers, I. and Sohn, C. "GAZE-2: An Attentive Video Conferencing System." In Extended Abstracts of ACM CHI 2002 Conference on Human Factors in Computing Systems. Minneapolis: ACM Press, 2002.