Media EyePliances: Using Eye Tracking For Remote Control Focus Selection of Appliances

Roel Vertegaal, Aadil Mamuji, Changuk Sohn and Daniel Cheng
Human Media Lab
Queen's University
Kingston, ON Canada K7L 3N6
{roel,mamuji,csohn,dc}@cs.queensu.ca

ABSTRACT
This paper discusses the use of eye contact sensing for focus selection operations in remote controlled media appliances. Focus selection with remote controls tends to be cumbersome as selection buttons place the remote in a device-specific modality. We addressed this issue with the design of Media EyePliances, home theatre appliances augmented with a digital eye contact sensor. An appliance is selected as the focus of remote commands by looking at its sensor. A central server subsequently routes all commands provided by remote, keyboard or voice input to the focus EyePliance. We discuss a calibration-free digital eye contact sensing technique that allows Media EyePliances to determine the user's point of gaze.

Author Keywords
Input Devices, Focus Selection, Eye Tracking, Attentive User Interfaces.

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

INTRODUCTION
With the continued convergence of wireless digital media devices and computing systems, boundaries between traditional media appliances such as TVs and stereos, and desktop computers have become blurred. Increasingly, we are moving to a world in which users interact with remote appliances that provide access to media sources that reside on networked computers. TiVo, Internet radio and AirTunes [1] are recent examples of this convergence trend. As a consequence, it seems, Human-Computer Interaction (HCI) is moving towards a multiparty human-computer dialogue, one in which computing applications are commanded remotely by a user through the embodiment of a corresponding appliance. In [2], Bellotti et al. posed a number of challenges for such multiparty HCI scenarios. In this paper, we hope to provide some suggestions towards answering two of these challenges: (1) How do I address one of many possible devices; and (2) How do I specify a target for my actions? How *do* we move from GUI-style interactions where multiple entities are represented on a single computing device to interactions with many remote devices in the real world? Appliance manufacturers have addressed such problems chiefly through the design of increasingly complex remote controls (see Figure 1). There are a number of problems associated with the use of such remotes. Firstly, unified remote controls have become complicated computing appliances by their own right, featuring a considerable learning curve. Almost all unified remotes now feature buttons to select a target appliance for their commands (See Figure 1). Pressing these focus buttons typically remaps the remote control configuration, effectively placing the remote control in a different appliance modality each time a focus button is pressed. This leads to considerable confusion for users alternating control between appliances.

Figure 1. Sony AV 3100 Universal Remote.

USING EYES FOR FOCUS SELECTION
Rather than using buttons, Shell et al. [10] proposed the use of eye gaze as a means for determining the target of user commands in situations with many appliances. There are a number of reasons why the use of eye gaze as a means for selecting a focus appliance is compelling:

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

1) In mobile scenarios, users do not need to carry an input device to perform basic pointing tasks. In scenarios where the hands are busy or otherwise unavailable, eye gaze provides an extra and independent channel of input.

2) The eyes have the fastest muscles in the human body, and consequently are capable of moving much quicker than any other body part. Moreover, researchers have reported that during target acquisition, users tend to look at a target *before* initiating manual action [6]. This means that eye gaze could provide one of the fastest possible input methods, if tracked effectively.

3) Users can produce thousands of eye movements without any apparent fatigue. Use of eye gaze mitigates the need for repetitive manual actions, and thus reduces the risk of repetitive strain injury.

4) Users are *very* familiar with the use of their eyes as a means for selecting the target of their commands. They use eye gaze during their communications with other humans to effectively indicate whom they are addressing or listening to [12]. Users are also familiar with others responding to them whenever they make eye contact [12].

However, it is important to distinguish between the use of the eyes as a continuous pointing device, and its use for selection of discrete targets. Indeed, users are *not* very familiar with the use of their eyes as a continuous pointing device. This is because the chief purpose of the eyes is to provide input to the human body, rather than provide output to control the exterior environment. Furthermore, the eyes do not typically perform well in continuous pointing. This is because the eyes move very rapidly between fixation points, to inhibit movement of the world on the retina [4].

A Midas Touch

There are other arguments against eye tracking as an input device. Firstly, the history of eye tracking has produced some grueling contraptions, chiefly aimed at keeping the user's head motionless. With advances in computer vision, however, we have now entered a realm where users can move relatively freely, with head movement tolerances in commercial trackers of over 30x15x20 cm. Secondly, eye tracking can be inaccurate and noisy. However, on-screen accuracies of better than 1 degree are now the norm. While this accuracy is still considerably less than that of manual pointing techniques, further improvements are likely. It has been suggested by Jacob [6] and others that the accuracy of eye trackers in pointing is fundamentally limited by the size of the human fovea, which is in the order of two degrees of visual angle [4]. According to this argument, there would be no need for the eye to position with greater accuracy than what is required to keep a visual target within the fovea. However, accuracy limitations are actually caused by current limitations in available computer vision algorithms [4]. Thirdly, eye trackers are still expensive. However, this is mostly caused by low market demand. Fourthly, until recently, eye trackers needed to be calibrated by having users track predefined targets. As we will demonstrate, discrete targets can now be tracked without any calibration whatsoever. Finally, eye trackers suffer from what is known as the Midas Touch Effect [6]. The Midas Touch Effect is caused by overloading the visual input function of the eye with a motor output task. It occurs chiefly when an eye tracker is used not only for pointing, but also for clicking. Providing clicks with the eyes is useful in cases where users do not have control over limbs other than their eyes. In such cases, the Midas Touch effect causes users to inadvertently select or activate any target they fixate upon. By issuing a click only when the user has fixated on a target for a certain amount of time (*dwell time click*), the Midas Touch effect can be controlled, but not entirely removed. The Midas Touch effect can be avoided by issuing clicks via an alternate input modality, such as a manual button or voice command.

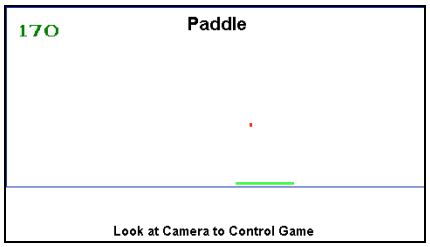

Figure 2. Eye-controlled Pong, with paddle tracking the horizontal coordinate of the eye.

Input = Output

More generally, if the output task interferes with the input task, the effectiveness of eye tracking input is greatly reduced. When mapping input to the eyes, it is therefore important to select cases where input equals output, i.e., where the movement of the eyes for the output task matches that required for visual input. A great example of such scenario is the use of an eye tracker to play Pong (see Figure 2). Here, as users are observing the ball, the horizontal coordinate of their eye movements is used to automatically move the paddle. The chief exploit of Pong, the eye-hand control problem, is thus mitigated, making eye-controlled Pong a game one cannot loose.

FOCUS SELECTION IN MEDIA APPLIANCES

Focus selection of remote controlled media appliances meets all of the above criteria for eye control. Firstly, most appliances provide a large and discrete target for eye fixations. Secondly, according to Wizard of Oz studies conducted by Maglio et al. [7] and by Oh et al. [8], users tend to look at a target appliance before issuing a command. While Maglio's study was limited to the use of voice commands, it has been suggested this behavior is also common in manual control [6].

Managing Media EyePliances

A Media EyePliance essentially consist of a normal home theatre appliance augmented with a miniature eye tracker, called an Eye Contact Sensor (ECS) [11]. Eye Contact Sensors allow Media EyePliances to sense when users are

Fig 3. EyeTuner: AirTunes Speaker with Digital ECS.

looking at them, without requiring any form of calibration. This allows users to determine which appliance is currently the target of remote control commands simply by looking at the appliance. Each Media EyePliance ECS is connected to a central server, which switches command between Media EyePliances upon user eye contact. When a Media EyePliance is used in conjunction with other Media EyePliances, this means commands are easily reused amongst devices. Commands can be issued by voice, remote control or Bluetooth keyboard. When a remote control is used, its commands are interpreted by the server and relayed to the appliance the user looks at through an RF, X10 or infrared transmitter interface. The chief advantage of this approach is that it allows users to control a large number of appliances without having to select from a large number of buttons, and without placing the remote control in a device-specific modality. In the case of voice recognition, the user need not carry an input device at all. Here, users can issue basic voice commands to a speech recognition engine located on the central server. Upon eye contact with a Media EyePliance, this speech recognition engine switches its lexicon to that of the focus EyePliance. After a command is interpreted, it is relayed to the appliance.

Calibration-Free Digital Eye Contact Sensors

Our eye contact sensors consist of an infrared camera, a set of infrared LEDs mounted on-axis with the camera lens, and a computer vision algorithm that runs on a networked computer system. The user's eyes are tracked by provoking a glint on the cornea using the on-axis light source. When the glint appears in the center of the onlooker's pupil, the user is looking at the sensor, and thus the EyePliance (see [11] for a more thorough discussion of ECS technology). The use of an ECS provides a cheap and easy method for EyePliances to track the user's gaze. However, there are a number of problems associated with the use of the methods proposed in Shell et al. [10,11]. Firstly, the resolution of analog eye contact sensors is generally insufficient for interactions at distances greater than approximately 120 cm from the EyePliance. To solve this problem, we developed a high-resolution 2MPixel digital version of the ECS. This digital ECS is capable of sensing eye contact by the user at 3 m. distance, with an angular resolution better than 9 degrees, and a head movement tolerance better than 1.5 m.

Resolving EyePliance Interference

When two Media EyePliances are placed within 70 degrees of visual angle from one another, the computer vision algorithm of either EyePliance may not be able to conclude which EyePliance the user looks at. This is because light sources interfere with one another. When the user is looking at EyePliance A, the glint produced by EyePliance A may in fact appear close to the pupil center, as seen from EyePliance B's perspective. To solve this problem, we designed a mechanism that allows multiple Media EyePliances to share eye contact sensing resources. A central server detects when two or more EyePliances simultaneously see the eyes of a user. When this occurs, it sends a signal to all but the most centrally located Master EyePliance. The peripheral EyePliances respond by turning off their ECS, instead flashing a pulse-code modulated ID code with their on-axis LEDs. This allows the Master to identify at which peripheral EyePliance the user is currently looking. The ECSes on the peripheral EyePliances begin functioning normally again when the user's eyes are lost for more than 6 seconds. The above procedure allows a large number of EyePliances to be placed within distances of approximately 50 cm of each other.

EyeTuner

Figure 3 shows the prototype of our first Media EyePliance, called *EyeTuner*. An EyeTuner essentially consists of a speaker with a digital ECS mounted on top, that allows the speaker to sense when users are looking at it. This speaker is connected over an AirTunes network to a computer running Apple's iTunes [1]. Whenever users produce a prolonged fixation at the speaker, our central server responds by lowering the volume of the currently playing song. If eye contact is sustained, it starts parsing user commands, whether issued by remote control, Bluetooth keyboard, or voice commands through a lapel microphone. Apart from recognizing standard remote control commands such as *play, pause* and *skip*, users can also query the iTunes library for tracks. Queries are performed using the Bluetooth keyboard, or via speech recognition. Users issue a speech query by saying "Find <*name*>" while looking at the speaker. Upon receiving the "Find" command, the speech engine switches its lexicon to the names of individual tracks, albums and artists within the user's iTunes library. The <*name*> query is subsequently submitted to iTunes over a TCP/IP connection. If a query results in multiple hits, EyeTuner responds by showing the list on its LCD display [9], after which it begins playing the first track. Users can subsequently skip through the tracks until the desired song is located.

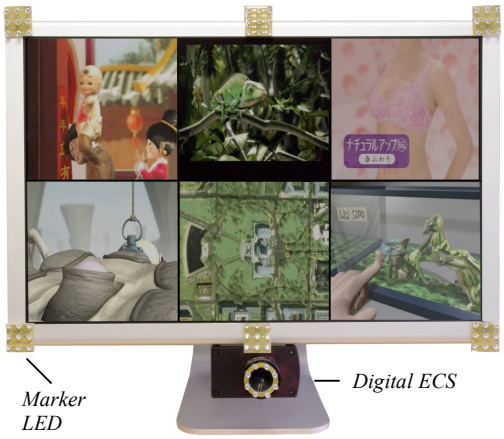

Figure 4. EyeDisplay with digital ECS and 6 Tiles.

EyeDisplay

Our second Media EyePliance is a 23" LCD television augmented with a digital eye contact sensor. Figure 4 shows how we augmented the television with 6 sets of infrared LED markers. These markers produce a corresponding reflection of a grid of glints on the cornea of the user. A digital infrared camera mounted below the screen tracks the eyes of the user, as well as the grid reflections that appear within those eyes. We utilized the fact that light sources *off-axis* to the camera appear in the center of the pupil when users look at them to develop a calibration-free eye tracking technique. When the user is looking at the top-right corner of the screen, the corresponding infrared marker appears centered in the pupil. When he looks at the bottom right of the screen that marker appears in the center of the pupil. A simple computer vision algorithm determines the point of gaze on the screen, with an accuracy of approximately half the distance between markers. This property is used to select channels, or media files residing on a networked media PC [5]. The display currently detects fixations at up to 6 tiles of media content at a time. Users select a media stream for playback by looking at its corresponding tile, and by pressing the *play* button on their remote control. Upon activation, the movie magnifies to fill the screen, and begins playing [3]. When the user hits the *stop* button on his remote, the screen goes back to its prior state. Similar to EyeTuner, the set of media files selected for display can be queried using a Bluetooth keyboard. Users can alternate use of their keyboard between Media EyePliances through looking, for example, to select music on EyeTuner to accompany a slideshow playing on EyeDisplay. Because the EyeDisplay TV watches the user, it can determine when it is being watched, and when not. When the user's eyes are lost, the currently playing media stream is automatically paused. When users resume watching the show, the stream automatically resumes play. Other uses of EyeDisplay include remote browsing of webcams, control of multiple remote computer systems, and instant messaging with multiple clients [5].

INITIAL EXPERIENCES

Initial evaluations of the use of eye input for focus selection proved encouraging. According to [5], focus selection with the eyes is about twice as fast as with hotkeys or mouse. Participants quickly learned to use their gaze for indicating the target of remote control commands. Users were able to switch control between EyeTuner and EyeDisplay with ease while playing and skipping media content. Users did note a lack of visual feedback on the detection of eye contact by an EyePliance. In response, we mounted a green LED on the ECS that turns on when the EyePliance gains focus. While we did not yet evaluate the use of voice control, switching of Bluetooth keyboard queries proved particularly promising. However, users complained about a lack of feedback on their keyboarding actions. This led us to display their keystrokes on the EyePliance display.

CONCLUSIONS

In this paper, we discussed the use of eye tracking as a means for focus selection operations in remote media appliances. We presented Media EyePliances, home theatre appliances that are augmented with a digital eye contact sensor. The eye contact sensor allows EyePliances to determine when the user looks at them. Eye contact is used to determine which EyePliance becomes the target of remote user commands. A central server automatically routes all commands provided by remote, keyboard or voice input to this focus EyePliance. The use of eye gaze allows the user to rapidly alternate control between EyePliances.

REFERENCES

1. Apple Computers, Inc. *www.apple.com/airtunes*, 2004.
2. Bellotti, V., et al. Making Sense of Sensing Systems. In Proc. of CHI 2002. Minneapolis: ACM Press, 2002, pp. 415-422.
3. Bolt, R. A. Gaze-Orchestrated Dynamic Windows. In Proc. of the 8th Annual Conference on Computer Graphics and Interactive Techniques, 1981, pp. 109-119.
4. Duchowski, A. *Eye Tracking Methodology: Theory & Practice*. Berlin: Springer-Verlag, 2003.
5. Fono, D. and Vertegaal, R. EyeWindows: Evaluation of Eye-controlled Zooming Windows for Focus Selection. In Proc. CHI'05. Portland: ACM Press, 2005 (in press).
6. Jacob, R.J.K. The Use of Eye Movements in Human-Computer Interaction Techniques. ACM Transactions on Information Systems 9 (3), 1991, pp. 152-169.
7. Maglio, P., et al. Gaze and Speech in Attentive User Interfaces. In Proceedings of the International Conference on Multimodal Interfaces. Berlin: Springer-Verlag, 2000.
8. Oh, A., et al. Evaluating Look-to-Talk: A Gaze-aware Interface in a Collaborative Environment. In Extended Abstracts of CHI 2002. Seattle: ACM, 2002, pp. 650-651.
9. Phidgets, Inc. LCD Display. http://www.phidgets.com, 2004.
10. Shell, J.S., Selker, T., and Vertegaal, R. Interacting with Groups of Computers. Communications of the ACM Vol. 46 No. 3 (March 2003), ACM Press, New York; 40-46.
11. Shell, J.S., et al. ECSGlasses and EyePliances: Using Attention to Open Sociable Windows of Interaction. In Proc.of ACM ETRA 04, San Antonio, TX, 2004.
12. Vertegaal, R., Slagter, R., Van der Veer, G., and Nijholt, A. Eye Gaze Patterns in Conversations: There is More to Conversational Agents than Meets the Eyes. In Proc. CHI 2001. Seattle: ACM Press, 2001, pp. 301-308.