What's in the EYES for Attentive Input

Eye-tracking systems hold some of the greatest potential among AUIs. Here, two systems that focus on eye gazing demonstrate how this simple form of visual attention can perform a level of common interactive tasks.

By Shumin Zhai

INPUT TO CONVENTIONAL COMPUTER SYSTEMS HAS BEEN limited to the user's explicit actions through the mouse and the keyboard. These input channels are rather impoverished in comparison to the increasingly rich multimedia output of computer systems, resulting in the asymmetrical bandwidth problem [3]. In such systems a single mouse-driven cursor conveys the user's attention, goal, action, and perhaps, even his or her frustration. This situation is likely to change in future Attentive User Interfaces (AUIs) with new channels of computer input based on computer vision and other technologies. AUIs may sense and track the user's presence, movement, body posture, head direction, and other environmental information (See the sidebar "How to Track What People Do"). A particularly rich source of information for AUIs is the user's visual attention as indicated by his or her eye gaze. Our hypothesis is that where the user looks contextualizes much of that user's action and hence can be applied to facilitate human-computer interaction (HCI).

MORE ON THE EYE

THE EYE IS BOTH A SERVOMECHANISM AND A MÉCANISME DE CERVEAU
AND SOMETIMES IT DOES ITS OWN THING AND SOMETIMES IT GOES WHERE THE BRAIN WANTS IT TO GO.
THE EYES ARE THE WINDOW TO THE MIND AND THE MIND'S WINDOW TO THE SCENE
SO THAT ONE IS NEVER QUITE SURE WHETHER IT'S THE WORLD OR THE MIND
THAT MAKES THE EYES SHIFT TO WHERE THEY'RE GOING FROM WHERE THEY'VE BEEN.
YOU CAN WATCH THE EYES AND SOMETIMES CATCH THE THOUGHT
WHILE IT'S SO HOT THAT THE MIND HASN'T HAD IT YET.
WITH A MIND OF ITS OWN THE EYE LOOKS AT THE PLACE BEST CALCULATED
TO LET THE MIND'S EYE SEE WHAT THE MIND WANTS TO SEE.
AND THEN ALL THE WORLD RUSHES IN TO BE REDUCED
TO COMMON SENSE AND PERCEPT BEFORE THE NEXT SACCADE IS LOOSED.

—JOHN W. SENDERS; "VISUAL SCANNING PROCESSES" (1983).

Opening more input channels to the computer brings us only halfway to actually enriching the user's computing experience. A critical challenge lies in integrating these new channels of information with the user's explicit input (see [1] for related challenges raised). We need to design unobtrusive, transparent, and subtle turn-taking processes that coordinate attentive input with the user's explicit input in order to contribute to the user's goal without the burden of explicit dialogues.

This article explores such mechanisms for exploiting one type of attentive input—visual attention as indicated by eye gaze—to perform a class of common interaction tasks: manual input control and target selection encountered in selecting menu items, entering and editing text, clicking Web links, and other low-level interaction activities.

Eyes Provide Context for Action

John Senders' poem (see the previous page) on eye movement elegantly and humorously captures both the potential value and possible pitfalls in using eye gaze as a source of HCI information. Because of the prominence of visual attention in human behavior, a great deal of user activity can be understood by observing the user's gaze. However, because eye movement can be involuntary, inferring the user's intention from eye gaze and applying the results to interaction is a tricky task.

An eye tracker (see the sidebar "Eye Detection with Dual Light Sources") can be used as an attentive input device that senses the characteristics of the user's eye movements, such as fixations, saccades (rapid intermittent eye movements between fixations), and history. We use eye tracking in combination with other sources of information, such as the current interaction task and the location of screen objects that might capture user attention, with the ultimate goal of integrating this information with manual control actions.

In mediating the use of eye-gaze information and user manual control action, we observe the following principles:

- The hands and eyes tend to work in combination.
- The eye gaze of a user tends to provide the context within which actions take place. As such, the use of the eyes for control input should be implicit.
- The hands of the user tend to act within the context of where the user looks. As such, the use of the hands for control input should be explicit.

Here, we describe two systems that illustrate how mechanisms based on these principles can be designed to seamlessly mediate an extremely subtle turn-taking process between the user and the attentive computer at almost subconscious levels of action and perception.

MAGIC Pointing

Eye control has long been a topic of interest in HCI (for example [2, 3, 5]), particularly for on-screen target pointing. Achieving "what you look at is what you get" without hand control clearly has much appeal. Apart from challenges in eye-tracking precision, there are two fundamental shortcomings to gaze-pointing techniques in which the entire control process is based on gaze tracking. First, pointing in HCI consists of two parts—moving a cursor to the target then activating the target (for example, with a mouse click). The second component is very difficult to handle with eye input alone. One obvious choice is to use eye blinks to click. The problem is we blink subconsciously most of the time. Another choice is to use gaze dwell time—setting a threshold of time duration beyond which the gazed object is activated. The

How to Track What People Do

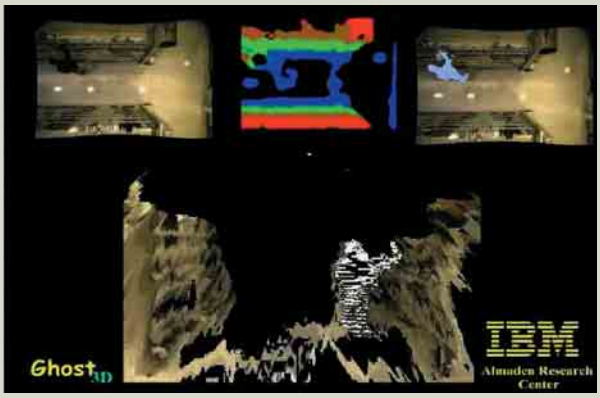


Figure 1a. Output of our tracking system.

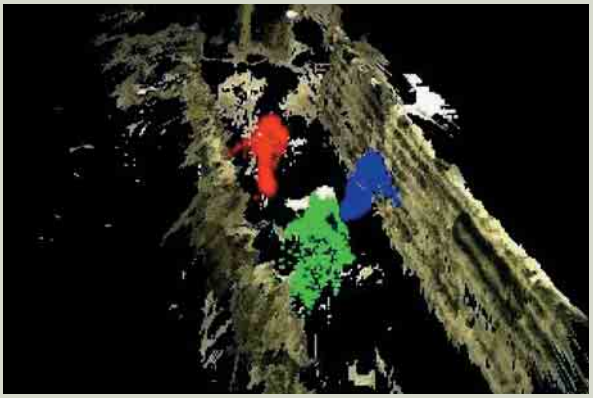


Figure 1b. Ghosts from the multiple cameras and multiple people.

By Ismail Haritaoglu, David Beymer, and Myron Flickner

By augmenting rooms with cameras, we can create an attentive environment that can automatically extract information about users' behaviors beyond their eye movements. We track multiple people using a network of ceiling-mounted stereo video cameras [1]. Our system outputs a time-dependant floor plan view of people in the space. At each time instance, each person's location along with their body posture, hand position, and head orientation are computed. Figure 1a shows a standard color image in the top left, a distance or range image computed from the stereo system in the top middle, and an example of a detected person's silhouette shown in light blue in the top right image. The silhouette is automatically extracted using both color and range background models. A 3D silhouette-*Ghost* is created using the 2D silhouette and the range data. The silhouette-*Ghost* is shown in white in lower image of Figure 1a. Using the silhouette-*Ghost,* the head and shoulder regions are identified using the constrained anatomical structure of the human body. By calibrating multiple cameras in 3D we can augment larger spaces and track multiple people as shown in Figure 1b.

An earlier version of this system was used to extract social shopping groups of people as they waited in a checkout line or service counter [2]. Once a person is detected, an appearance model based on color and edge density in conjunction with a mean-shift tracker is used to recover the person's trajectory. People are grouped together as a shopping group by analyzing inter-body distances.

New data collection techniques using intelligent stereo cameras, along with sophisticated database mining algorithms, will transform the study of human behavior, enabling the world to be a laboratory for human behavior analysis. C

References

1. Haritaoglu, I., Beymer, D., and Flickner, M. Ghost3D: Detecting body posture and parts using stereo. In *Proceedings of the IEEE Workshop on Motion and Video Computing* (Orlando, FL, Dec. 2002).
2. Haritaoglu, I. and Flickner, M. Detection and tracking of shopping groups in stores. *IEEE International Conference on Computer Vision and Pattern Recognition* (2001).

problem with this method is that users have to be conscious of where and for how long they are gazing, in order to avoid triggering actions unintentionally.

Second and more generally, the use of eye gaze in humans to perform control tasks does not tend to occur naturally. The eye, as a primary perceptual organ, does not function very well as a control organ. As mentioned earlier, eye movements are not always under voluntary control. To load the visual perception channel with a motor-control task seems fundamentally at odds with natural behavior in which the eye searches for and takes in information from the world and in which the hand manipulates objects in that world. Other than for disabled users, the direct use of eye gaze for pointing does not appear very promising.

Given the natural division of function between the eyes and the hand, it is therefore important to use eye gaze as an implicit source of information about the user's visual attention during tasks such as pointing. In this approach, instead of controlling the cursor location directly with the eyes, we dynamically optimize the cursor's start position according to the user's gaze at the start of pointing. We then let the user manually guide the cursor to its final destination. Such an attentive pointing system would contain the following subsystems (see Figure 1):

- Track the user's eye gaze in relation to the computer screen using computer vision. Separate moments where the eye fixates (fixations) from moments where the eye moves (saccades). Estimate current location of the user's gaze on a screen.
- Track and analyze interactive context, including the state of the user's activity based on interaction history, location of user interface objects, and their probabilities of being the next target (hot spots).
- Observe, manage, and control the pointing process, including identification of the states of a pointing trial, particularly the beginning of a pointing gesture. The system places the cursor at the estimated current gaze point, which is typically on or near a target. If a hot spot with high target probability is identified in the vicinity of the estimated gaze point, the system should place the cursor on the spot.

Through this process, the system tries to optimize pointing by gauging the object of the user's attention based on eye gaze and modeling the available attention attractors on the screen.

Once the cursor's start position is defined, the user needs only to make minor adjustments of the cursor using a manual input device to reach the target. This is called Manual and Gaze Input Cascaded pointing (MAGIC, also for Manual Acquisition with Gaze Initiated Cursor, see Figure 2.) Because the user typically looks at a target on the screen to acquire it before pointing, it seems as if the MAGIC pointing cursor simply appears where the user's task needs it to be. A critical property of MAGIC pointing is the user need not be aware of the role of eye gaze, since it is used only implicitly.

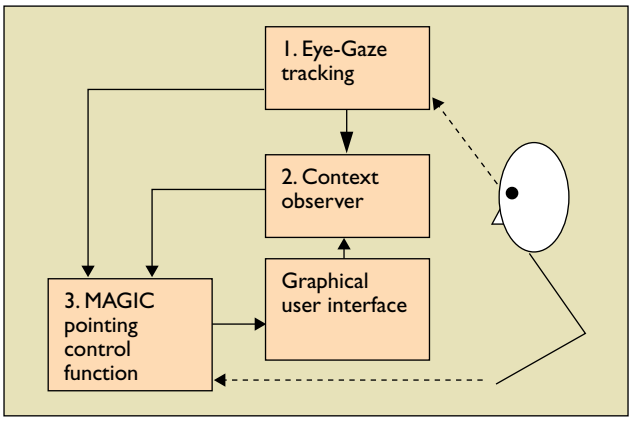


Figure 1. Gaze-attentive system for manual control.

To mediate between eye-gaze-based input and the user's manual control, we designed both liberal and conservative approaches to trigger the MAGIC cursor. In the liberal approach, the cursor (new) home position is defined by every new object at which the user looks. A new object is defined by its distance (for example, 120 pixels) from the current cursor position. The liberal approach may appear proactive, as the cursor waits readily on or in the vicinity of every potential target. However, such an overactive cursor may be distracting to the user, and can be detrimental to his or her attentive focus. In the future, it is conceivable the system could include functions that decide if a new cursor should be placed according to the context of interactions. If the user is reading text, for example, a MAGIC cursor should not be moved.

The more conservative MAGIC pointing technique does not place a cursor at a target until the manual input device has been actuated. To minimize directional uncertainty, the cursor home position is offset in the opposite direction of the hand stroke (or force vector in case of a pointing force stick) rather than placed at the center of the gaze area. This means that once warped, the cursor is likely to appear in motion toward the target, regardless of the direction in which the user actuated the manual input device. This conservative approach would never be overactive since it does not allow the cursor to jump to a target the user does not intend to acquire. However, it may be slower than the liberal approach.

The first advantage of MAGIC pointing is the user can select objects of arbitrary size, which traditional eye-gaze control cannot, using only a short cursor movement and a button press. MAGIC pointing is therefore well suited for small, finger-operated input devices such as touchpads or the pointing sticks embedded in keyboards, since both can be easily optimized for small precise movement. Second, the user never loses the cursor—the cursor appears in the fovea of the user's eye with every new pointing gesture. Furthermore, it does not impose any additional burden or fatigue on the eye since eye-gaze information is only used implicitly and unconsciously. The eye must look at the target of interest with or without MAGIC pointing.

To test the feasibility of MAGIC pointing, we implemented an experimental system using an eye tracker developed in our laboratory based on a dual infrared illumination method (see the "Eye Detection with Dual Light Source" sidebar). We conducted an experiment [6] in which the two MAGIC-pointing techniques were tested against a pure manual pointing method. The results show that both MAGIC-pointing techniques are promising alternatives to manual pointing. All participants completed the tasks successfully. On average, completion time was 1.4 seconds with the standard manual control technique, 1.52 seconds with conservative MAGIC pointing, and 1.33 seconds with liberal MAGIC pointing. Although by the end of the experiment, participants had fewer than 10 minutes of exposure with each of the two novel MAGIC tech-

niques, they were able to use these at a speed similar to that of well-practiced manual control on average, but with less manual work. The conservative MAGIC technique showed a trend of further improvement during the course of the experiment. Furthermore, the best trial performances in the experiment were consistently achieved with the MAGIC-pointing techniques, showing their potential if eye-tracking technologies were faster and more stable.

Our experimental eye tracker operated at only 30 frames per second. Since capturing a fixation point may take more than two samples, the eye-tracking speed in the experiment may have limited the advantage of MAGIC. With a higher-speed eye tracker (for example, 100Hz), more speed advantage is expected from MAGIC pointing.

Subjectively, participants enjoyed the fact that the MAGIC-pointing system attended to their needs by placing the cursor at or near where it should be. Some participants were clearly disappointed when moving from a MAGIC pointing session to a purely manual session as they realized the cursor would no longer magically appear in the vicinity of the target.

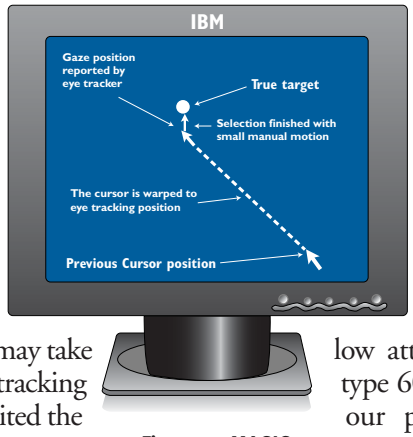


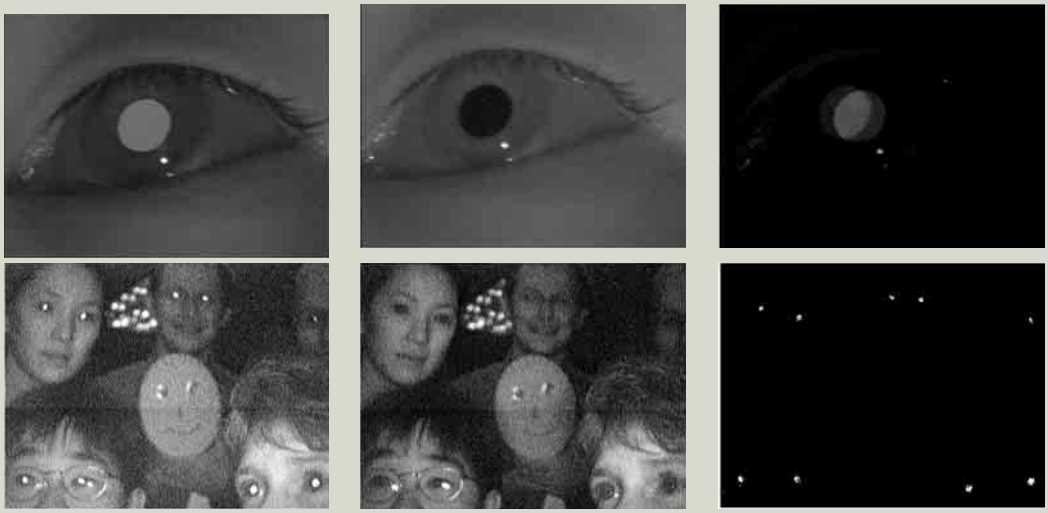
Figure 2. MAGIC pointing on screen.

EASE Chinese Input

Text entry is another one of the most frequent HCI tasks. Although progress has been made toward speech and handwriting recognition, typewriting remains and will likely be the main text-entry method in the future. Once learned, touch-typing has two critical advantages: rapid speed and low attention demand. A skilled typist can type 60 words per minute, much faster than our potential for handwriting. Another, more important property is that touch-typing frees the user's visual attention so users can focus their attention completely on the visual display. It is interesting to study how attentive input devices such as eye trackers may aid this process.

Numerous difficulties arise when the user's language

Eye Detection with Dual Light Sources



By Myron Flickner, David Koons, and Carlos Morimoto

Before you can track gaze you must first find the eyes in the image. One simple way to do this is to exploit the "red eye" effect commonly seen in flash photography. Most mammalian eyes are retro-reflectors—light from a source will be returned in the direction of the source. Changing the geometric relationship of the camera and the light source can intentionally induce or avoid bright pupils, or red eyes. The IBM PupilCam [2] uses two light sources synchronized to the video fields. The bright pupil image is subtracted from the dark pupil image and adaptively thresholded to create a computationally simple eye detector [1]. C

References

1. Morimoto, C.H., Koons, D., Amir, A., and Flickner, M. Pupil detection and tracking using multiple light sources. *Image and Vision Computing 18* (2000), 331–335.
2. Koons, D. PupilCam Construction Instructions. IBM Research Report RJ 10212, (2001); www.research.ibm.com/resources/paper_search.shtml.

is Chinese (or many other nonalphabetic languages). Currently, the most popular method used for Chinese text input in Mainland China is pinyin input. Pinyin is the official Chinese phonetic alphabet based on Roman characters. The complication of pinyin input, however, is that most Chinese characters are homophonic. In Mandarin Chinese, each syllable corresponds to 16.8 Chinese characters on average. As a result, when a user types the pinyin of a character, the computer software displays many candidate characters with the same pronunciation, together with identifying numbers, typically in a one-line graphical window. For example, the first eight candidate characters for "zhong" could be

1中 2种 3重 4众 5钟 6终 7忠 8肿

which corresponds to eight different meanings. The user must then select one of the multiple candidates by typing the identifier number, such as "1." Often, this requires the user to look away from the screen in order to identify the location of the number on their keyboard. Furthermore, when the candidates are displayed on more than one line, the user must scroll to another line of candidates by hitting a page key.

We designed and implemented a prototype system called Eye Assisted Selection and Entry (EASE) for Chinese text input [4] based on an approach similar to MAGIC's—implicit use of eye input for determining the possible focus of user attention, rather than for direct control, and use of this attentive input to introduce a seamless, subconscious turn-taking process between the user and the computer. With EASE, the user types pinyin as usual but presses the spacebar as soon as he or she sees the intended target character in the list. Because the system attends to the user's eye gaze, it correctly selects the intended character (the one closest to the user's current gaze location) when the space bar is pressed.

A user study showed people successfully used the EASE system for Chinese typing [4]. Without practice, users could complete their typing task with EASE as fast as the conventional method. More importantly, the study shows that users could select the intended target character with EASE faster than with the traditional numeric keying method. The study also showed that EASE allowed users to employ their visual attention on the screen more optimally by eliminating the need to look at their numeric keys, thus reducing the cognitive load associated with this task. EASE is a clear example of how an AUI may help optimize low-level skill-based routines and visual attention within the confines of a traditional GUI.

Conclusion

How to effectively combine attentive input with more explicit user input actions is a critical challenge in our research toward AUIs. MAGIC and EASE demonstrate it is possible to find appropriate places within interaction to integrate the two input streams for each specific task, even at the tightly closed-loop perceptual motor level of HCI. In both cases, eye gaze was used to provide the possible context of user attention, which was implicitly combined with the user's manual action. Thus, MAGIC and EASE take advantage of the user's natural behavior—paying attention to the target object of interaction before manipulating it. Because of the dominance of the hand in current interface control tasks, we have discussed manual input only, but the same principle also applies to commands issued by foot, lip, and tongue movement or speech. MAGIC and EASE demonstrate the great potential of AUIs in optimizing such tasks.

In prototyping and experimenting with MAGIC and EASE we experienced some of the limitations of current eye-tracking technologies that may conceal the potential of these novel methods. Efforts are being made by researchers and companies all over the world to improve the speed, reliability, and range of eye-tracking devices. Finally, it is important to note there is more to attentive input than meets the eye. In an effort toward ubiquitous computer vision of overall behavior of the user's body, we have implemented a tracking system that measures user presence, position, orientation, and movement path (see the "How to Track What People Do" sidebar). Given the rapid progress in camera- and image-processing hardware and in computer vision algorithms and software, we foresee a time when attentive computing devices that utilize information of the eyes, head, and body of users will become commonplace. C

References

1. Bellotti, V.M.E., Back, M.J., Edwards, W.K., Grinter, R.E., Lopes, C.V. and Henderson, A. Making sense of sensing systems: Five questions for designers and researchers. In *Proceedings of ACM CHI02* (Minneapolis, Apr. 2002), 415–422.
2. Bolt, R.A. Eyes at the interface. In *Proceedings of Human Factors in Computer Systems* (1982). ACM Press, NY, 360–362.
3. Jacob, R.J.K. What you look at is what you get: Eye movement-based interaction techniques. In *Proceedings of ACM CHI90.* Addison-Wesley/ACM Press, NY, 11–18.
4. Wang, J., Zhai, S., and Su, H. Chinese input with keyboard and eye tracking—An anatomical study. In *Proceedings of ACM CHI01* (Seattle, Apr. 02), 349–356.
5. Ware, C. and Mikaelian, H.H. An evaluation of an eye tracker as a device for computer input. In *Proceedings of ACM CHI+GI.* (1987) ACM Press, NY, 183–188.
6. Zhai, S., Morimoto, C., and Ihde, S. Manual and gaze Input cascaded (MAGIC) pointing. In *Proceedings of ACM CHI99* (The Hauge, The Netherlands, Apr., 1999). ACM Press, NY, 246–253.

Shumin Zhai (zhai@almaden.ibm.com) is a research staff member at the IBM Almaden Research Center, San Jose, CA.

CONTRIBUTING TO THIS WORK WERE CARLOS MORIMOTO, STEVE IHDE, JINGTAO WANG, HUI SU, BARTON SMITH, PAUL MAGLIO, AND MYRON FLICKNER.