A GAZE-RESPONSIVE SELF-DISCLOSING DISPLAY

India Starker† and Richard A. Bolt

The Media Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

india@dandelion.ci.com
bolt@media-lab.media.mit.edu

ABSTRACT

An information display system is described which uses eye-tracking to monitor user looking about its graphics screen. The system analyzes the user's patterns of eye movements and fixations in real-time to make inferences about what item or collection of items shown holds most relative interest for the user. Material thus identified is zoomed-in for a closer look, and described in more detail via synthesized speech.

KEYWORDS: Eye tracking, self-disclosing systems

INTRODUCTION

While the use of eyetracking in conjunction with computer-based information systems has yet been relatively rare, there have been a few instances where feedback from tracking the user's point-of-regard on a display has been used in real-time to modulate a display. These include: controlling stimulus presentation in the conduct of perceptual studies [21]; dynamic alteration of presented sentences as a function of where the observer is looking [12]; insuring no stimulus to peripheral vision areas by presenting test stimuli only where the subject is looking, or only when the subject is looking in the desired spot - a king of controlled "tunnel vision" [8].

There are also examples from rehabilitation engineering, where eyetracking has enabled the severely disabled who have only the use of their eyes to gain some degree of control over their environment: to pick out messages on a display screen [5,17]; to pick out codes to operate appliances, such as videotape players [13,20]; even to steer electric-powered wheel chairs, the rider looking in the direction they wish to go [16].

Tracking an observer's eye has been used to control a dynamic display of many video episodes [4], and, on a limited-bandwidth display, to devote sharpest resolution to where the observer is looking, that is, at the point of the observer's sharpest (foveal) vision [3], and in flight simulators to project a high-resolution insert in line with the pilot's direct line-of-sight, wherever on the display that may be, against a less-detailed background scene [6].

The eye is in fact an excellent "pointer" [19], and the eye has been found to be a serviceable in "picking out" items on a CRT display [22]. In the applications cited above, the display reacts more or less directly to the user's momentary point-of-regard used as a kind of one-on-one pointer. In contrast, in the work reported here, the system operates more *interpretively*, aggregating series of eye fixation points over the immediate past to evaluate the *pattern* of fixations so sampled as evidence upon which to make inferences about what material currently on display holds most interest for the user. Items and areas thus determined to be of high relative interest are then shown off more closely and described in more detail in synthesized speech.

EYES AND INTEREST

A person's eye movements and eye fixations correlate strongly with a person's interest in, and attention to, things in their surroundings [8,9]. People tend to look at what attracts them, especially at what they find curious, novel, or unanticipated [1,2,11]. In particular, Yarbus [23] reported that an observer's pattern of looking at a scene is systematically influenced by their specific interests as manipulated by questions put to them about that scene. This finding particularly encourages the notion that eye movements and fixations can be valid clues to inferring an observer's interest.

†Present address: Cognition Corporation, 755 Middlesex Turnpike, Billerica, Massachusetts 01821

True, one can be attending to something, yet not looking directly at it [7,15]; and one can look at something yet not be attending to it. Mainly, however, when we are in fact paying attention to things in our visual surround, the eye's point-of-regard is a very good index of the distribution of that attention [9, pp. 50-65]. Such observations, both from everyday life and supported by controlled experiment, suggest that a computer system that can detect where on its graphics display the user is looking can use that information to infer the pattern of the user's interest in the things on display and to help formulate the elaboration of that information in both visuals and synthesized speech narration.

The behavior of such a system would be akin to that of the skilled salesperson or interviewer who pays close attention to the "body language" of another person, in particular to where the eyes are trained, as cues to the other's inner state. The aim, in this instance, would be to make reasonable inferences about what items or cluster of items currently on display hold most interest for the user/observer. Having determined the item(s) or region in which the user seems most interested by aggregating the locus of eye fixations over the very recent past, the system would then proceed to "zoom in" upon the item or items and tell more about it or them in synthesized speech. Narration about the item or items of interest continues until: a) the system exhausts its store of things to say about the item or items; b) the system judges that the user, on evidence of looking patterns, seems interested in something else (coherent eye patterns centering somewhere else) or possibly satiated with the current subject (looking that is relatively uncorrelated with what is presented). The overall result, ideally, is that the system is "self-disclosing" in that it shows off its database according to the interests exhibited in user eye actions, and at a pace that matches the user's own.

THE SYSTEM

In our set-up, the observer is seated before the 19-inch diagonal color monitor of a Hewlett Packard Series 9000/835 workstation with a Turbo/SRX graphics accelerator board. This machine is specialized for computer graphics, with matrix multiplication conducted in hardware.

The observer's eye is about 20 inches from the monitor's screen. Eye tracking is done with an ISCAN Pupil/Corneal Reflection Tracking System. In this system, a low-intensity infra-red lamp illuminates the user's eye. A video camera, sensitive in the infra-red range and positioned near the light source, is focused upon the eye. By means of video image processing, the ISCAN system determines the darkest spot--the center of the pupil--and the brightest spot--the reflection in the cornea of the lamp filament--at 60 Hz or about every 16 milliseconds. (Cf. Figure 1.) These values are passed to an IBM XT personal computer which in turn calculates the distances in X and Y between pupil center and the corneal reflection, which distances vary systematically as the observer's eye looks about, independent of small movements of the head. The observer's head is steadied in a chin-rest so as not to go outside the view of the camera. Lastly, a Digital Equipment Corporation DECTalk speech synthesizer provides speech output.

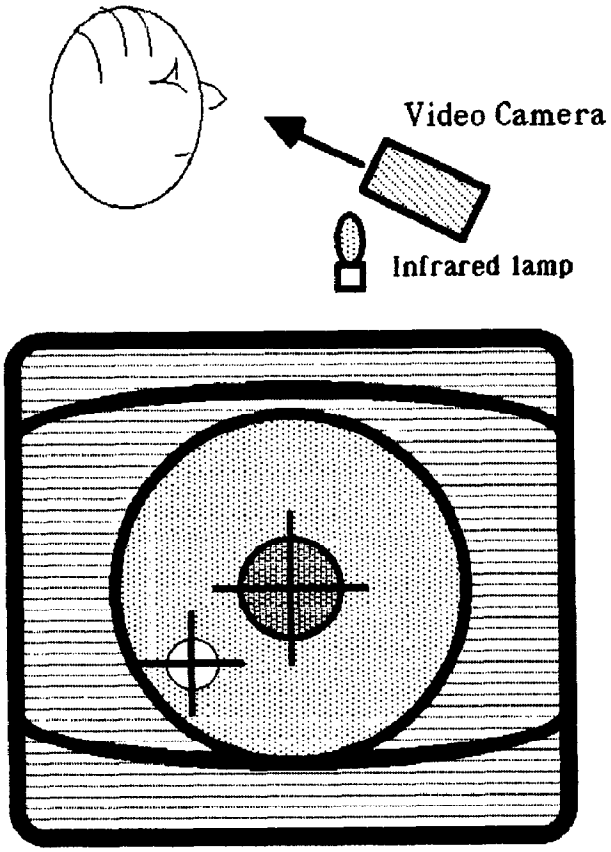

Figure 1.

In the pupil-center/corneal-distance method of eyetracking, the centers of the pupil and of the reflection in the pupil of a reference dot of light (lamp filament) are determined, and the distance between these two centers is measured. This distance is linearly related to changes in the observer's point-of-regard, independent of small movements of the head.

Antoine de Saint Exupery's book *The Little Prince*, provided the inspiration for a whimsical 3-D graphics world to be shown to our subject observer/users: a small planet slowly revolving agains a background of twinkling stars (Figure 2-a). The planet bears a number of features: volcanos on its surface; small staircases; flowers in orbit. The source of the synthesized speech narration could of course be disembodied, but we chose instead to have a visible narrator. Thus, a two-dimensional face of the "Little Prince" is shown in the upper left corner of the screen.

THE SYSTEM IN ACTION

Should the user be glancing about the displayed scene generally, the synthesized speech coming as it were from the depicted Little Prince tells about the planet as a whole, e.g., "This is where I live. It's not very large, maybe about two hundred feet across, but I call it home...". Should the user's looking at a pair of the staircases, shifting glances back and forth between the two, the plausible inference is that they are interested in the staircases as a group rather than in one specifically. In that case, the system's commentary might run: "My latest hobby is collecting staircases. I find it more relaxing than collecting stamps..." or some such commentary about the staircases as a pair or group.

However, should the user be found to be focusing for the most part on one staircase rather than the pair on view, the commentary might run "I found this green staircase on a trip by Pluto..," that is, providing comments specific to a certain item and only that item. Thus, the generality or specificity of the system's narration is a function of the scope and focus of the user's attention, whether wide and covering a group of items or indeed the entire scene, or focused in upon some single thing, as inferred from the user's pattern of eye fixations.

While we chose a 3-D graphics world with an item in motion (the revolving planet), the graphics scene could be as well static and/or planar. In any event, our planet revolved slowly enough so that the observer could see things without strain; and, on most occasions, we in fact "stopped" the world and simply showed the planet motionless. Figure 2-b shows the eye path tracing of a student subject observer looking about the temporarily stopped planet.

Note that this eye path tracing shows concentrated looking upon the planet's two staircases with an excursion over to the area where the face of the Prince appears. In this sampling, looking was concentrated first upon one staircase and then upon the other *sequentially*. Thus, in this case, narration was not about the staircases as a pair, but was first about the rightmost staircase, and then about the one on the upper left edge of the planet's image.

The system's narration at any level about some particular thing or group of things--or the world as a whole if looking were justly dispersed about the entire world scene--would continue until: a) the system ran out of things to say, and would simply say so and stop talking, at least about that thing; b) the person began to look elsewhere, including breaking into patterns of looking that did not correlate in any particular way with the features of the world-view being shown.

MODEL OF USER INTEREST

In the system, an "interest module" would find the object(s) currently being looked at, and record the time. (Screen coordinates of the gaze point were associated with three-dimensional objects by use of a ray-casting and bounding sphere test. Hidden objects--on the other side of turning planet--were discarded using geometrical constraints.) A looked-at object was caused to "blush," as feedback to the observer, by momentarily lightening its color on the display.

It was important that the recorded time an object was seen be as close to the actual time as possible. The time was recorded whenever an x-y or *screen coordinate* was read off the port to the IBM XT where the eye-tracking calibration code was resident. This screen coordinate was associated with an object or objects, and the object was time-stamped with the just recorded time.

Three models were implemented for determining the apparent level of interest of the user in a given object:

Model One: When the screen coordinate of the gaze point corresponds to an object or objects, the tally for that object is incremented by one. The interest level equals the tally.

Model Two: The elapsed time since a given object was seen is multiplied by a constant, k2, and subtracted from a constant, k1, times the tally of glances for that object:

$$interestlevel = k1 * tally - k2 * elapsedtime$$

Model Three: Model three, with its decay model of memory, may be the most realistic. In this model, whenever there is a fresh look at an object, the old value is decayed by the proper amount and then incremented by a constant (FreshLook Constant):

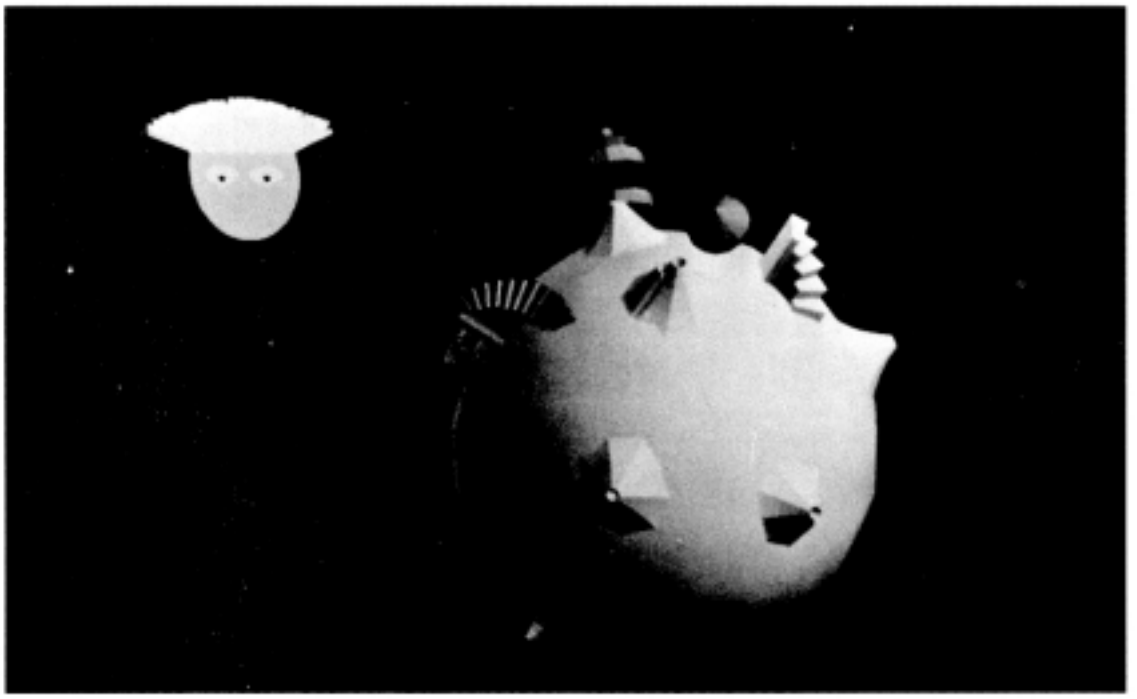

Figure 2-a.

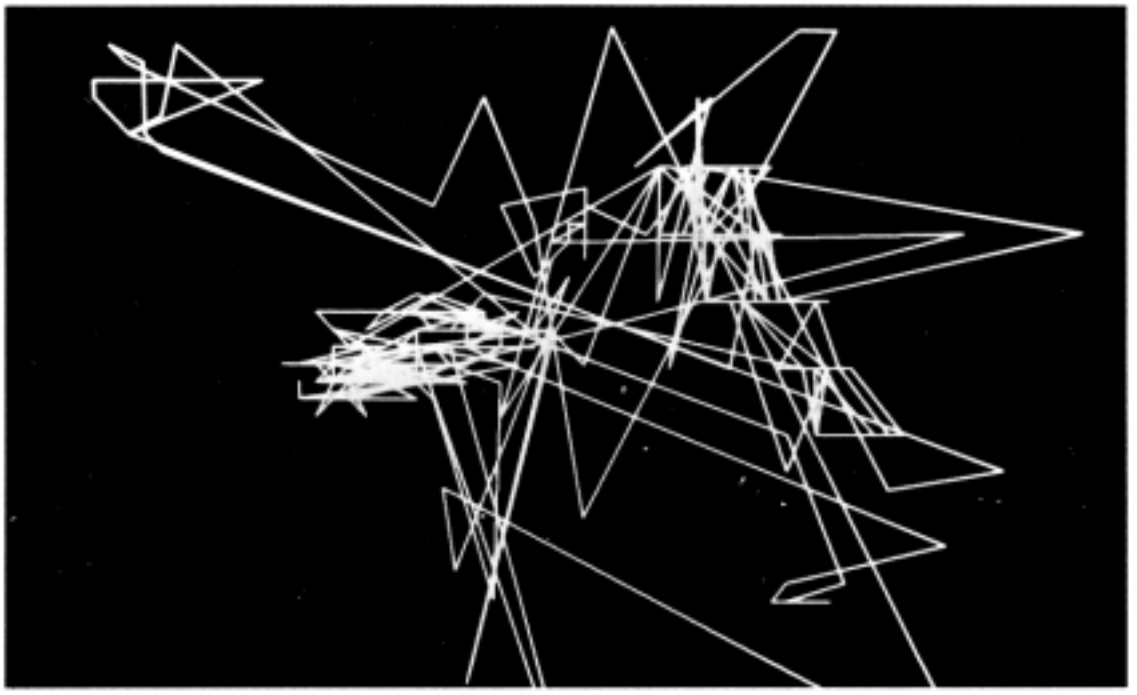

Figure 2-b.

if (object was just looked at)

$$interestlevel=(interestlevel)\cdot e^{\frac{-t}{\tau}}+FreshLookConstant$$

else

$$interestlevel=(interestlevel)\cdot e^{-\frac{t}{\tau}}$$

where

FreshLookConstant = constant
t= elapsed time since object was last seen
τ= time constant

For Model One only, the slate is wiped clean--all interest levels are reset to zero when a subject is chosen. The other two models inherently drop the interest level variable with elapsed time, accomplishing a similar end, but more gracefully.

DECIDING WHAT'S OF INTEREST

The general schema of the programmed "story teller" is:

- interest-stamp all objects
- find object(s) with highest interest level
- determine next subject
- talk

After a pre-set time for data collection (usually set for 2.0 seconds), the interest levels for all the objects are computed, and the object with the greatest interest level is found. The object, however, is not necessarily the subject for the synthesized narration. The next subject for narration is found from the user interest level of all the objects.

If there is a group of objects with interest levels near the one with the highest value, these objects are compared to see whether or not they belong to the same group of category (e.g., both are staircases). For example, in Figure 3, if interest level is plotted for each object, objects 2 and 5 might fall into a cluster. A sample standard deviation is used to determine this cluster of *similar* interest levels--all the interest levels are tested to see whether or not they fall within some fraction of one standard deviation.

In our prototype system, with its simple world, three degrees of "generality" are stipulated, the minimum number to illustrate the possibility of having a cluster of like objects qualify as the subject of narrative. With a more elaborate graphics scene and database, one may have many levels of generality.

In our system, the degree of generality is determined from the cluster information. If the cluster contains objects of the same category, for example, objects 2 and 5 in Figure 3 are both volcanos, the middle degree of generality--objects of same type, here volcanos--is chosen; otherwise, the most general degree (the whole planet) is chosen. If there is a single object that stands out by itself in the cluster information, the most specific degree is chosen: the narration is about that specific object. As a result, the subject talked about by the Prince in our system is always one of the following:

1) an object, spoken about specifically
2) similar objects, spoken about as a category
3) the world in general

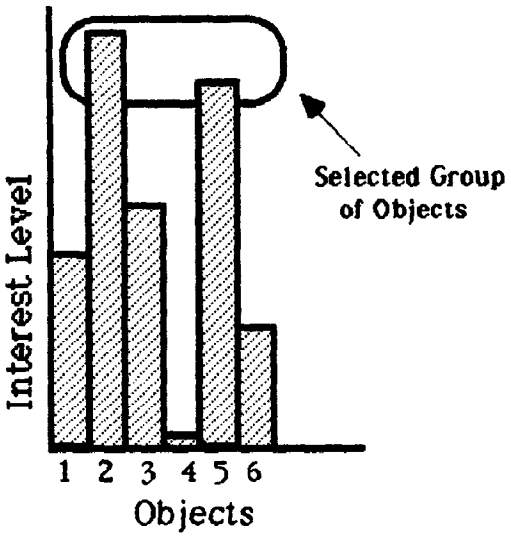

Figure 3.

PERSONALITY VARIABLES

The apparent personality of the narrator (here, the "Prince") can be controlled by the constants used in the computation of interest levels.

For instance, shortening the amount of time the user is observed before a narrative topic is chosen causes continual speech and immediate response to a glance. The narrative, however, quickly falls behind the user, if this observation time is too short. Experimenting with briefer phrases as narrative content might offset this effect.

The narrator will continue to speak about a selected subject in models one and three, above, when there is

no new object being looked at. The point at which the narrator quits *persisting* on the same topic is manipulated by a low-end cut-off point of the sample standard deviation of item interest levels. (A variable, *"wishy-washy,"* sets how long the narrator will talk boldly on about the same thing.) Even though the values of the interest levels fall with time, the same object will continue to have the greatest value. In the instance where a group of items had been looked at, the standard deviation falls as well, with the effect that the same group is chosen again. Therefore, until a new object is seen, which spreads the sample standard deviation, the same subject is chosen. It is not desirable that the narrator continue talking about an object if the user/observer is no longer looking at the object(s). This is managed by changing the action once the standard deviation gets very small.

In Model 3, manipulating τ has the effect of lengthening or shortening the narrator's memory: an increased τ causes a longer memory, and vice-versa.

DISCUSSION

Samples of looking were gathered from student subject observers, and were stored for playback as desired. This playback feature of the system proved to be very useful in comparing the behavior of the models on identical input. Additionally, performance of the three models was compared using sequences of simulated input where the path of the eye was stipulated by inputing values "by hand," in our case via a knob-controlled eye cursor. These looking sequences could be exactly controlled, to be played back for each of the models.

Differences in narrating behavior showed up between the three models. For instance, with a sample of simulated input which "looked" first at a volcano on the planet and then on a staircase, it was found that a difference between models one and three lie in that, in model three, the volcano's high interest level faded out slowly enough for it to continue to be selected at a point in time when model one would have started to talk about the staircase. Model 3, with the same input tended to speak longer on a subject than the other two; model three takes longer, generally, to reach the low-end threshold of "interest" that effectively cuts off a subject. Such differences are not, of course, surprising since different calculations are performed in the models to produce variants in "reactivity" and "persistence" of the narrator [18].

In the literature on gaze, the *duration* of a glance is considered an important quantifier [10]. The models used here do not incorporate the concept of duration as such [18]. In further work, the creation of a "duration" variable out of consecutive fixations on the same object might be explored.

Possible elaborations to the "story teller" or the virtual personality that provides the narrative include: transitional text, to "bridge" across interruptions and returns to any subject, e.g., (upon coming back to a previous topic) "...As I was saying about the blue staircase..."; having the narrator recognize when he is being looked at, with appropriate spoken responses to reflect this fact; animating the narrator's graphic mouth to synchronize with speech synthesizer output.

Potential elaborations to the general logic of the system include attempting by closer examination of observer looking patterns to distinguish "casual" from "intense" interest, and to adjust system reactivity and pace of narration accordingly.. Also, the narrative itself might be made more sophisticated by paying more attention to discourse strategies [14]. As indicated above, the current system does not "know" what it has previously said; sentences are not repeated merely because the text file that keeps them is not rewound. A more sophisticated approach wherein the system contemplates narrative content might make the narrator all the more convincing to the user/observer.

ACKNOWLEDGEMENTS

The authors wish to thank Dave Berger, Edward Herranz, David Koons, and James Puccio for their able assistance and valuable contributions to this research.

Support for this work was provided by Grants IRI-8615741, and IRI-8746936 from the U.S. National Science Foundation, Division of Information, Robotics, and Intelligent Systems.

REFERENCES

1. Berlyne, D. E. Curiosity and explanation. *Science, 153* (1966), 25-33.
2. Berlyne, D. E. The influence of complexity and novelty in visual figures on orienting responses. *Journal of Experimental Psychology, 55* (1958), 289-296.
3. Bolt, R. A. *The human interface.* Van Nostrand Reinhold, New York, 1984.
4. Bolt, R. A. Gaze-orchestrated dynamic windows. In *Proceedings of ACM SIGGRAPH Computer Graphics Conference* (Dallas, Texas, August 3-7). ACM, New York, 1981, pp. 109-119.
5. Chandler, D. L. "Breaking the shackles of disability." *Boston Globe* newspaper, Boston MA,

Technology Section, February 6, 1989, p. 31-32.

6. Elmer-DeWitt, P. Into the wild blue (digital) yonder. *TIME*, August 1, 1988, 62-3.

7. Jonides, J. Voluntary versus automatic control over the mind's eye's movement. In John Long and Alan Baddeley (Eds.), *Attention and Performance IX.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1981.

8. Just, M. A. and P. A. Carpenter. The role of eye-fixation research in cognitive psychology. *Behavior Research Methods and Instrumentation, 8*, 2 (1976), 139-143.

9. Kahneman, D.. *Attention and effort.* Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

10. Kleinke, C. L. Gaze and eye contact: a research review. *Psychological Bulletin, 100*, 1 (1986), 78-100.

11. Loftus, G. R. and N. H. Macworth. Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology, 4*, 4 (1978), 565-572.

12. McConkie, G. W., D. Zola, G. S. Wolverton, and D. D. Burns. Eye movement contingent display control in studying reading. *Behavior Research Methods & Instrumentation, 10*, 2 (1978), 154-166.

13. McGowan, J. The eyegaze computer. *Occupational Therapy Forum* (Atlantic Edition), Vol. III, No. 47, week of November 21, 1988, 1-5.

14. McKeown, K. R. Discourse strategies for generating natural-language text. In Barbara J. Grosz, Karen Speck Jones, and Bonnie Lynn Webber (eds.), *Readings in Natural Language Processing.* Morgan Kaufmann publishers, Inc., Los Altos, CA, 1986. 479-500.

15. Posner, Michael I. Orienting of attention. *Quarterly Journal of Experimental Psychology, 32* (1980), 3-25.

16. Rinard, G. A. and E. E. Rugg. Current state of development and testing of an ocular control device. *Proceedings of 4th Conference on Systems and Devices for the Disabled,* Seattle, Washington, June 1977.

17. Rogers, M. "More than wheelchairs." *Newsweek* magazine, April 24, 1989, p. 66-67.

18. Starker, B. E. I. *A Gaze-Directed Graphics World with Synthesized Narration.* Unpublished master's thesis, Massachusetts Institute of Technology, 1989.

19. Steinman, R. Role of eyemovements in maintaining a phenomenally clear and stable world. In Richard A. Monty and John W. Senders (Eds.), *Eyemovements and psychological processes.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1976.

20. Tello, E. R. Between man and machine. *BYTE, 13*, 9 (September, 1988), 288-293.

21. Vaughn, J. Control of visual fixations in visual search. In: Senders, John W., Dennis F. Fisher, and Richard A. Monty (eds.), *Eye Movements and the Higher Psychological Functions.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1978.

22. Ware, C. and H. H. Mikaelian. An evaluation of an eye tracker as a device for computer input. In *Proceedings of CHI + GI '87 Human Factors in Computing Systems* (Toronto, Canada, April 5-9, 1987), ACM press, pp. 183-188.

23. Yarbus, A. L. *Eyemovements and vision.* Translated by B. Haigh. Plenum Press, New York, 1967.