Towards gaze-mediated interaction: Collecting solutions of the “Midas touch problem”

Boris Velichkovsky, Andreas Sprenger and Pieter Unema

Applied Cognitive Research
Dresden University of Technology
D-01062 Dresden, Germany
Fax +49 (351) 4637741
E-mail velich@psy1.psych.tu-dresden.de

ABSTRACT: For a development of truly user-centered interfaces we need to take into account not only generic characteristics of human beings but also actual dynamics of attention and intentions of persons involved in an interaction. Modern eyetracking methods are indispensable tools in such a development, as they allow the use of eye movement data for control of output devices, for gaze-contingent image processing and for desambiguation of verbal as well as nonverbal information. The main obstacle on the way to these applications is the so-called “Midas touch problem”: how to differentiate “attentive” saccades with intended goal of communication from the lower level eye movements that are just random or provoked by external stimulation? We report results of our investigations of the problem and present a solution based on a functional classification of fixations correlated with their duration. Several additional solutions are also considered together with the data on the trainability of the human oculomotor system.

KEYWORDS: Attention, eyetracking, fixations, levels of processing, user-centered approach

INTRODUCTION

Technological changes often grow up silently in interdisciplinary neighborhoods. New methods from cognitive neurosciences may well soon initiate such changes in human-computer interaction. For the first time, neuroimaging and eyetracking methods provide evidence about the actual course of human information processing and brain structures involved (Velichkovsky & Hansen, 1996). This is exactly the type information we would need to adapt our interfaces to momentary fluctuation in abilities, performance and intentions of users. In particular, the new eyetracking methods have reached the threshold of ergonomic acceptance for the alternative forms of interaction (Hansen, Andersen & Roed, 1995, Jacob, 1995) and are considered in

professional publications as one of the most promising technologies for the development of cognitive ergonomics in the short and middle term perspective (Joch, 1996). In our contribution we consider some empirical questions which have to be answered on the way to a broad practical use of these methods.

EYETRACKING AND HUMAN-COMPUTER INTERACTION

The new generation of eyetracking systems do afford different applications with a relatively low load for their users. They are increasingly fast, precise and non-intrusive. As a rule, the systems are video-based, which means that the gaze direction is computed on the basis

Figure 1: Human eye in infrared light of an eyetracking camera. Right: A non-intrusive eyetracker at the work place.

of picture of the pupil and the corneal reflex from infrared light sources. In the non-intrusive versions the only demand on the user is a more or less constant sitting position whereby the head movements are usually permitted (see Figure 1). The precision of the systems is a 0.5 degree of arc and is independent from blinks that are identified as such by software. Temporal resolution (50 or 60 Hz) however is too low for recognition of short fixations (Starker & Bolt, 1990, Velichkovsky, Hillers & Sprenger, in preparation). Head-mounted systems can be much faster, up to 600 Hz, and are even more precise in terms of spatial resolution, especially with the interfaces based on the self-organized neural networks (Ritter, 1993) that lead to a further improvement of resolution up to 30%. (Baluja & Pomerleau, 1994, Pomplun, Velichkovsky & Ritter 1994). The precision of about 0.5 degree of arc seems to be sufficient for most of the applications. During the session users only notice the light head-mounted device. Their head movements and slight body movements are permitted. Modern eyetracking systems are built according to the TCP/IP standard, so a transmission of data via Internet is possible (Velichkovsky & Hansen, 1996).

Depending on the system one can measure different parameters, such as blinks, pupil's diameter and, last but not least, the gaze direction and the fixation time. These parameters are interesting primarily as indices of the focus of visual attention of the users. Indeed, the gaze direction is considered to be the most reliable, though unfortunately not ideal (Helmholtz, 1896, Hoffman, 1996), index of the focus of attention.

Combining of all these parameters one can find indices for the type of processing and the human emotional state (Velichkovsky, 1995). The most obvious application of the methods is the "gaze-mouse", as e.g. in situations where it is no longer possible to rely on the fine manual coordinations or on the speech control of the interface, but the gaze direction can be *interpreted* in an intentional way (Stampe & Reingold, 1994, Starker & Bolt, 1990). A more sophisticated approach to gaze-based interaction is less connected with the idea of using gaze as a substitute for the computer mouse. Instead, the data are to be *interpreted* at a higher level for a reconstruction of user's perception (Pomplun, Ritter & Velichkovsky, 1996, Raeithel & Velichkovsky, 1995) or to be used in a new type of "noncommand interface" (e.g., Nielsen, 1993, Velichkovsky & Hansen, 1996). In other words, eye movements are used as a continuous measure of attention being paid to individual objects on the display. Even more complex *interpretation* of the data may be needed for development in the direction of "interest and emotion sensitive" media (Engell-Nielsen et al., 1997 in press).

We are now coming to the main issue. As already mentioned the gaze direction is not an ideal indicator of attention: e.g., we cannot jump to point B while attending point A, but we can well attend B while fixating A. During an average working day our eyes make more than 100.000 saccadic movements. Which of them is attentive or initiated with a communicative intention and which is not? How should computers decide? No wonder that some of the previous attempts to use eyetracking data for an interacting with a

computer have not been particularly successful (Starker & Bolt, 1990). Jacob (1995) pointedly termed this difficulty the "Midas touch problem". Obviously, one should try to filter out the tremendous amount of data, but according to what criteria and for what tasks?

A MODEL OF VISUAL INFORMATION PROCESSING

To answer these questions we need a model of eye movement control in relation to information processing tasks. In the last time this was the discovery of *express saccades* that raised an interest to eye movement data (e.g. Fischer & Weber, 1993). The movements have an extraordinary short latency of less than 100 ms, the shortest latency of any movements in humans. Functionally they seem to belong to the organism's orienting response: they are in a sense reflexes imposed on us by our environment and they resist any conscious control. This means that their potential importance for the interaction with computer is "zero".

A classification of the higher order tasks for eye movements still is a somewhat speculative matter (see however Challis, Velichkovsky & Craik, 1996, Deacon, 1996, Velichkovsky, 1994). First of all, a very basic function of eye movements (which is performed in a synergy with other motor systems of the organism) is *search*. If a potentially important object or feature has been discovered it should be perceptually described. Such an *identification* is hardly possible without eye movements, though these movements can be of another type than during the search. After identification the information can be semantically categorized. Some of the information may receive also additional *self-referential* (or perhaps *other-referential*) interpretation. Only at this point the information may be considered to be ready for a possible *communication*. That also at this last stage eye movements are important has been claimed more than 100 years ago (Hering, 1977), but remained almost unexplored. In a series of experiments we recently investigated parameters of eye movements in all these tasks.

Figure 2: Typical pattern of saccadic eye movements in a search task

Eye movements in search tasks

A typical pattern of search saccades is shown in Figure 2 (Velichkovsky, Challis & Pomplun, 1995; Pomplun et al., 1997 submitted). The task was in detection of one single difference between left and right parts in the display of objects which varied in color and form. One can easily see the high amplitude saccades scanning the picture in the way that an immediate repetition of the visit to the same region is being blocked. There is a neurophysiological mechanism behind this variant of the “travelling salesmen problem” known as the IOR (Inhibition of return). It inhibits after a fixation (probably only in the search modus) all the saccades directed to this particular region of the space (Hoffmann, 1996).

The fixation time found in these experiments usually lies between 120 and 250 ms. The time may be longer in the vicinity of the critical elements of the display shortly before an answer. This probably demonstrates an interference from the higher levels because repeated fixations of the same region are possible in the final phase. From the point of view of possible applications, this all means that the diapason of fixation times immediately above the region of express saccades, until at least 250 ms, has to be regularly closed for a command initiation in gaze-mediated interfaces.

Levels of processing and fixation time

If a potential source of information is found it can be processed so-to-say to different “depths”: either one remains “on the surface” of encoding of physical features or one tries to reconstruct semantic and personally relevant relations of the data as well. The actual level of processing is a function of the situation, the task and the user’s skills and abilities. The investigation of the levels is a traditional chapter of cognitive studies, whereby one usually differentiates at least three levels, that of perceptual, semantic and self-referential processing (Challis, Velichkovsky & Craik, 1996, Velichkovsky 1994, Velichkovsky, Challis & Pomplun, 1995). There is recent evidence from investigation in our lab that eye movements correlate with the level of the encoding tasks. Figure 3 shows an example of material from some of these studies (Velichkovsky, Hillers & Sprenger, in preparation) that allows a control of levels of processing and the difficulty of perception.

A systematic effect observed in such levels of processing experiments is that in the case of perceptual encoding (i.e. encoding oriented to physical features of objects) there are relatively many fixations with duration between 120 and 250 ms. With a deeper encoding – semantic or self-referential – the number of fixations within the segment of 250-450 ms grows significantly. What do these data tell us? Obviously even after time of fixation of 250 ms there has to be some further internal information processing aimed at a reconstruction of semantics and personal relevance of the situation. Before this higher-order information is extracted it is simply too early to use the gaze position for communication. In what segment of fixation time are there traces of our communicative intentions?

Are there specific “communicative fixations”?

To take the gaze direction into account for communicative purpose is a natural human ability. Early in the development of the face-to-face interaction we learn to coordinate our attentional resources with

Figure 3: Stimulus material from the levels of processing experiments (three grades of sensory reduction is shown)

those of other people, so that the state of “joint attention” becomes possible: we attend the same object at the same time as our partner and the problem of the high referential ambiguity of natural language disappears (Velichkovsky, Pomplun & Rieser, 1996). The lacking support for the “joint attention” states may be one of the reason for the ergonomic problems of electronic communication, especially in transmission of spatial and practical knowledge. In Figure 4 one can see the layout of one of our recent experiments aimed at investigation of these issues (Velichkovsky, 1995.).

The task consisted in a cooperative solution of a puzzle game. One of the two subjects was an expert, who solved the task many times before, the other – a novice, who saw the puzzle for the first time but should find a solution in practice. We registered verbal communication of the partner and their eye movements in the following three modes of communication:

- “Voice”: expert and novice could communicate only by voice.
- “Voice and mouse”: in addition to verbal remarks it was possible for the expert to show relevant parts of puzzle with the mouse.
- “Voice and gaze”: in addition to verbal communication actual position of fixations (either of expert or of novice) were shown on the monitor of the partner.

For the present context the data on the eye movements of partners are of special interest. An explicit intention to support verbal instructions by the gaze leads to a significant rise of the proportion of

extremely long fixations – around 500 ms and more. There is also a negative correlation of the number of such fixations in experts and the puzzle solution time in novices. In other words, the long fixation can be indeed used as a source of information in a situation of an indirect communication with other people mediated by computer.

CONCLUSIONS AND TASKS FOR THE FUTURE

With respect to the questions formulated at the beginning of the paper these experiments give us a rather straightforward answer. Different segments of fixation time may have a different functional interpretation. In the order of tasks from the initial detection and search to a more “in-depth” processing leading to understanding of the information and its communication to partner in cooperative problem solving, there is a systematic increase in the duration of fixations. Particularly, in the case of communicative intentions one finds extremely long fixation of 500 ms and more. A simple visualization of the results is presented in Figure 5. Additional neurophysiological studies recently demonstrated that these changes in levels of processing also involve different brain mechanisms (Challis, Velichkovsky & Craik, 1996, Velichkovsky et al. 1996). This corresponds to the known data about hierarchical organisation of the control mechanisms of eye movements (Fischer & Weber, 1993).

The practical solution to the “Midas touch problem” is therefore to set up the temporal filter for the “gaze-mouse” applications on the values of 450-500 ms. The threshold seems to be of importance also for other reasons. In the course of verbal description of a scene (as well as in reading), eyes are often a little bit ahead of speech, so that there is a temporal gap between voice and eyes – Eye-voice span. The span is variable, but as a rule it is about 400-500 ms or less (Velichkovsky, Pomplun & Rieser, 1996). In other words, only when the gaze position will be explicated with a latency of about 500 ms, the intended focus of speech and the actual focus of visual attention will converge on the same object. Such a redundancy supports the verbal instructions and creates a context for the “joint attention” states.

Figure 4: Experiments on the transfer of gaze direction data in cooperative problem solving

Figure 5: Fixation time as an indicator of the level of processing

There are further technological options of updating the “gaze-mouse” to a completely reliable communication tool. The “gaze-mouse” does not need a visual feedback in contrast to the usual, manually controlled computer mouse. However such a feedback (e.g. in the form of “eyecons” – a dynamic visualization of eye closing and initiating commands, if fixated during the time of more than 500 ms) seems to be helpful from the point of view of users (Hansen, Andersen & Roed, 1995, Velichkovsky & Hansen, 1996). A later development can be an eclectic combination of gaze direction, fixation time, pupil diameter and possibly blinking frequency as a measure of interest or different interpretation of the display recognized by an individually tuned neural network (Nielsen, 1993, Ritter, 1993, Velichkovsky & Hansen, 1996).

With this possibility we would bridge the gap between the traditional “gaze-mouse” story and the second, more advanced group of applications based on a continuous monitoring in distribution of attention. In this case it is the set of fixations over the time which defines the course of interaction with computer. A

particularly interesting technique is the method of “attentional landscapes” illustrated by the Figure 6. Here two more traditional ways of representing the eye movement data (as fixation points (a) and as clusters of fixations (b)) are confronted with the “attentional landscape” – a more graded representation of the resolution of details based on the idea of an idealized fovea (as in (c)). The “attentional landscape” can be used as an image filter function (see Figure 6(d)) to reconstruct the subjective view of the initial picture, so-to-say “as seen by the user’s cortical homunculus”.

We were able to show in recent time that the use of attentional landscapes as a filter function for the gaze-contingent processing of visual displays does allow to adequately visualize the actual subjective perception, e.g., of ambiguous pictures (Pomplun, Ritter & Velichkovsky, 1996, Velichkovsky, Pomplun & Rieser, 1996). With the data summarized in the present paper we can in the future improve the interface design by visualizing of a “first look” of an interface or by reconstructing – in the form of images – the history of user’s attempts to go beyond the information given, as in efforts to understand semantics of one or another

Figure 6: Different ways of representing eye movement data: (a) as points or circles, (b) as clusters, (c) as “attentional landscapes”, (d) as subjective view of the picture (after Velichkovsky, Pomplun & Rieser, 1996))

element of the display. The knowledge about functional significance of different classes of eye movements is an indispensable step on the way to this goal, because the selection of eye movement parameters for gaze-mediated interaction has to be task-dependent. What we cannot say at present is how eyetracking may change the parameters of human eye movements. According to some reports (Cleveland & Cleveland, 1996) a systematic use of “gaze-mouse” devices permanently improves user’s performance, so that even thresholds as low as 250 ms can be finally managed. Perhaps our eyes can in a sense repeat the evolution of the hand – to an instrument of new, specifically human skills shaped by the development of technology.

ACKNOWLEDGMENTS

We appreciate support of our colleagues and co-workers in Dresden, Bielefeld and Toronto. Thanks are also due to Nancy and Dixon Cleveland (LC Technologies, Inc., Fairfax, VA) for their data on trainability of human oculomotor system.

REFERENCES

- Baluja S. and Pomerleau D.(1994). Non-intrusive gaze-tracking using artificial neural networks. In: *Neural information processing systems, Vol. 6*, New York: Morgan Kaufman Publishers.

- Challis, B.H., Velichkovsky, B.M. and Craik, F.I.M. (1996). Levels-of-processing effects on a variety of memory tasks: New findings and theoretical implications. *Consciousness & Cognition*, 5, 63-89.
- Cleveland, D. and Cleveland, N. (1996 September) Personal communication.
- Deacon, T. W. (1996). Prefrontal cortex and symbolic learning. In B.M.Velichkovsky and D.M.Rumbaugh (Eds.), *Communicating meaning*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Engell-Nielsen, T., Glenstrup, A.J. and Hansen, J.P. (1997 in press). Eye-gaze interaction: A new media – not just a fast mouse. *International Journal of Human-Computer Interaction Studies*,
- Fischer, B. and Weber, H. (1993). Express saccades and visual attention. *Behavioral and Brain Sciences*, 16, 553-610.
- Hansen, J.P., Andersen, A.W. and Roed, P. (1995). Eye-gaze control of multimedia systems. In Y.Anzai, K.Ogawa and H.Mori (Eds), *Symbiosis of human and artifact*. Amsterdam: Elsevier Science Publisher.
- Helmholtz, H.v. (1896). *Handbuch der physiologischen Optik*, 3d edition. Hamburg und Leipzig: L.Voss.
- Hering, E. (1977). *The theory of binocular vision*. New York: Plenum Press [first German edition, 1868].
- Hoffmann, J.E. (1996). Visual attention and eye movements. In H.Pashler (Ed.), *Attention*. London: University College London Press.
- Jacob, R.J.K. (1995). Eye tracking in advanced interface design. In W.Barfield and T.Furness (Eds.), *Advanced interface design and virtual environments*. New York: Oxford University Press.
- Joch, A. (1996). What pupils teach computers. *Byte*, 21, 99-102.
- Nielsen, J. (1993). Noncommand user interfaces. *Communications of the ACM*, 36(4), 83-99.
- Pomplun, M., Ritter, H. and Velichkovsky, B.M. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception*, 25(8), 931-948.
- Pomplun, M., Velichkovsky, B.M. and Ritter, H. (1994). An artificial neural network for high precision eye movement tracking. In B.Nebel & L.Drescher-Fischer (Eds.), *Lectures notes in artificial intelligence*. Springer Verlag, Berlin.
- Pomplun, M. et al. (1997 submitted). Comparative visual search. *Journal of Experimental Psychology: Human Perception & Performance*.
- Raeithel, A. and Velichkovsky, B.M. (1995). Joint attention and co-construction of tasks. B.Nardi (Ed.), *Context and consciousness: Activity theory and human-computer interaction*. Cambridge MA: MIT Press.
- Ritter, H. (1993). Parametrized self-organizing maps. In S.Gielen and B.Kappen (Eds.), *ICANN93-Proceedings*, Berlin: Springer.
- Stampe, D.M. and Reingold, E. (1994). Eye movement as a response modality in psychological research. In *Proceedings of the 7th European conference on eye movements*, Durham, University of Durham,.
- Starker, I. and Bolt, R. A. (1990). A gaze-responsive self-disclosing display. In *CHI 90 Proceedings*. New York: ACM Press.
- Velichkovsky, B.M. (1994). The levels endeavour in psychology and cognitive science. In P.Bertelson, P.Eelen and G.d'Ydewalle (Eds.), *International perspectives on psychological sciences: Leading themes*. Howe UK: Lawrence Erlbaum Associates.
- Velichkovsky, B.M. (1995). Communicating attention: Gaze position transfer in cooperative problem solving. *Pragmatics and Cognition*, 3, 199-222. 23.
- Velichkovsky, B.M., Challis, B.H. & Pomplun, M. (1995). Working memory and work with memory. *Zeitschrift für experimentelle Psychologie*, 42, 672-701.
- Velichkovsky, B.M. and Hansen, J.P. (1996). New technological windows into mind. In: *CHI 96 Proceedings: Human factors in computing systems*. New York: ACM Press.
- Velichkovsky, B.M., Hillers, R. and Sprenger, A. (1996 in preparation). Encoding, eye movements and memory performance in the processing of non-verbal information.
- Velichkovsky, B.M., Pomplun, M. and Rieser, H. (1996). Attention and communication. In W.H.Zangemeister et al. (Eds.), *Visual attention and cognition*. Amsterdam: Elsevier Science Publisher.
- Velichkovsky, B.M. et al. (1996). Evoked coherences of EEG: Communication of brain regions and levels of processing. *Zeitschrift für Elektroenzephalographie, Elektromyographie und verwandte Gebiete*, 27, 111-118.