Enhanced Gaze Interaction Using Simple Head Gestures

Oleg Špakov
University of Tampere
Kanslerinrinne 1, 33014, Tampere, Finland
oleg.spakov@uta.fi

Päivi Majaranta
University of Tampere
Kanslerinrinne 1, 33014, Tampere, Finland
paivi.majaranta@uta.fi

ABSTRACT

We propose a combination of gaze pointing and head gestures for enhanced hands-free interaction. Instead of the traditional dwell-time selection method, we experimented with five simple head gestures: nodding, turning left/right, and tilting left/right. The gestures were detected from the eye-tracking data by a range-based algorithm, which was found accurate enough in recognizing nodding and left-directed gestures. The gaze estimation accuracy did not noticeably suffer from the quick head motions. Participants pointed to nodding as the best gesture for occasional selections tasks and rated the other gestures as promising methods for navigation (turning) and functional mode switching (tilting). In general, dwell time works well for repeated tasks such as eye typing. However, considering multimodal games or transient interactions in pervasive and mobile environments, we believe a combination of gaze and head interaction could potentially provide a natural and more accurate interaction method.

Author Keywords

Eye tracking, head gestures, dwell time, selection.

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces – Evaluation/methodology; Input devices and strategies

General Terms

Experimentation, Human Factors, Performance.

INTRODUCTION

Dwell-time is usually recognized as the best method for making selections in gaze-controlled interfaces (e.g. [7, 21]) and remains the most popular selection method, especially in interfaces targeted at people with disabilities [2, 16]. In literature, a wide range of applications are based on dwell time and fixation driven selection, such as typing (e.g., [15]), drawing [7], chess playing [22], and many others.

UbiComp' 12, Sep 5 – Sep 8, 2012, Pittsburgh, USA.

The dwell-selection method is often criticized due to the high concentration that is required from the users. First, the user has to fixate on the item for long enough for it to be selected, which can be tiresome and slow down the interaction. On the other hand, the user must be careful *not* to stare at the selectable objects for longer than the predefined time interval, as doing so causes undesired selections or double entry errors (e.g., [9, 23]). There is a long history of search for replacements for the dwell time and solutions for the inherent Midas touch problem [25].

Discrete and continuous eye gestures have been found useful as an alternative selection method. In certain cases, gestures seem to provide better interaction than dwell time. Continuous gesturing (smooth pursuit driven selection) is mostly used for eye typing [5, 24, 26, 27] where it can significantly speed up the typing of consecutive characters. However, it is supposed to be problematic or even impossible to use in other context. Discrete (finite) gestures (saccade-driven selection) using various targets – off-screen [10, 11], explicit/strong [1, 27] or implicit/weak [3, 6] – can be used to complete a wide range of tasks (e.g., [18]). However, such gestures often result in slower interaction (compared to dwell time), they may not be convenient to use, and they may cause higher error rates for some users [6, 28].

There are also other known eye-based alternative selection methods but many of them either have obvious disadvantages or they lack experimental evidence to prove them as truly useful alternatives for dwell time. For example, voluntary blinks [20] are difficult to distinguish from natural blinks and thus require an extended blink (kind of dwelling) which not only slows down the interaction but also blocks the user's sight for that time. For a review of other alternative selection methods for gaze based interaction, see e.g. [8] or [21].

The most useful alternative selection methods are usually based on other modalities than gaze, leading to multimodal interaction. Kumar et al. [13] provide a good example of such multimodal interaction. In their study, gaze was used only for pointing and the selection was triggered by a key-up event. In other studies, frowning [23] and other facial expressions [4] were used as a signal to trigger selections while pointing with gaze. The problem which these multimodal interfaces is, that they often require several devices and systems, usually one device (system) per modality, for example, an eye tracker for gaze detection and an electromyography (EMG) biofeedback device for detecting facial muscle activity. Thus, in addition to the eye tracking device, the user may need to wear other specialized devices.

We propose using a single eye-tracking device for hands-free interfaces with input from gaze and head. Most of the commercial (e.g., LC Technologies EyeGaze, Tobii Technology 50-and T- series) and open-source (e.g., ITU Gaze Tracker) systems provide the eye position in the camera view (EP_{CW}) in real time, therefore head gestures can be detected by tracking the EP_{CW} (validated in several studies, e.g., [12]). We propose using this data for detecting voluntary head gestures, which could potentially replace dwell time for selection. Furthermore, different types of head gestures could to be assigned to a set of functions, such as navigation (e.g., turning and scrolling pages, changing a level in hierarchical structures, etc.) and functional context (mode) switching. Mode switching is needed when similar gaze events are used for different purposes, for example, a glance on the left may cause different actions depending on the currently active mode or context [11].

The design and development of head gesture recognition algorithms is currently a popular topic in the field of video-based interaction. The focus in these articles is on the technology itself, but user studies are still rare. The same applies for the few studies that deal with HCI systems that are used for detecting both the head gestures and the eye movements [14, 19]. The work that is most close to ours was recently published by Mardanbegi et al. [17] on combined use of eye pointing and head gestures with a head mounted tracker. We used a remote tracker; thus, the algorithms and data handlings are different. We also developed our own method for gesture detection, described later in this paper.

We will first list the primitive gestures selected to be tested, followed by the description of the detection algorithm used to recognize gestures, and finally present and discuss the results of this preliminary study. We were especially interested in 1) evaluating the applicability of a simple non-learning head gesture detector, 2) the stability of gaze detection in frequent and fast head motion, 3) subjective evaluation of the users' fatigue caused by frequent head gestures, and 4) subjective evaluation of the gestures by participants who are experts in using dwell-time for interaction with gaze-controlled interfaces.

HEAD GESTURES DETECTION

Five simple head gestures were selected for testing: nod, turn left, turn right, and tilt left, tilt right. We hypothesized that a quick nod could be the best head gesture for selection (most frequent operation). Turning the head is a natural way to point the direction and it also resembles the hand gesture when turning pages, therefore this gesture could be used for navigation. Finally, tilting as the slowest among these gestures could be used for mode switching (least frequent operation). More complex gestures consisting of these primitives were left out of the current study, to be tested in future experiments.

Previous work on head gesture recognition using eye detection mostly relies on Hidden Markov Models (HMM) for creating the detectors (e.g., [12]). HMMs can be quite robust in distinguishing between three head states (neutral, nodding and shaking), with up to 97% of correct detections, as presented in [12]. However, in other studies [19, 24] where the number of states was higher, HMMs were found notably less robust (80-90%).

Due to the fact that we tested several gestures in this experiment, and for the sake of simplicity in development, we used a range-based algorithm (RBA) which searches in EP_{CW} patterns of movements with direction, amplitude and duration falling into predefined ranges. The pattern of each gesture can be described by a few simple stages that the EP_{CW} should pass through sequentially. By "stage" we mean a time interval during which EP_{CW} moving parameters (speed and direction) remain about constant. For example, a nod gesture can be treated as consisting of four stages: 1) the EP_{CW} is relatively stable, 2) the EP_{CW} of each eye moves down, 3) the EP_{CW} of each eye moves up, 4) the EP_{CW} is relatively stable. The first and the last stages are common for all gestures and used as indicators of the gesture's temporal and spatial limits. The second and the third stages are distinct for each gesture, and can be expressed by the range of two EP_{CW} movement properties – angle and amplitude – within a certain time interval, as shown in Figure 1.

In our experiment, we used an eye-tracking device that reports the EP_{CW} in a range between (0;0) and (1;1). Five detectors (one per gesture) with four predefined stages for each were used in analyzing the data in parallel in real time. The RBA ranges of each stage and each detector are shown in Table 1 (valid when sitting vertically at ~60cm distance from the device and the user's eyes are aligned with the screen upper edge). The values of these ranges were estimated during the analysis of gestures recorded in a pilot test with two participants. The ranges selected were those that resulted in the correct detection of all gestures made during the pilot tests.

It is important to note that the eye movement itself only causes very small changes of the EP_{CW}. Thus it is impossible to simulate, for example, a nodding head gesture simply by glancing down and up. Uncontrolled body and

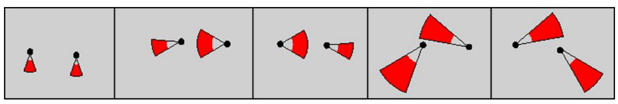

Figure 1. The eyes (black circles) in the camera view and the areas of each gesture (red) to move onto by the end of the second gesture stage (screenshots from a training session).

Stage	Gesture	Eye	Amplitude x10^{-2}		Angle, °		Interval, ms	
			min	max	min	max	min	max
1, 4	all	both	0	0.5	0	360	80	120
2	nod	both	1.5	4.0	250	290	100	200
	shake left	left	3.0	6.0	175	210	200	350
		right			150	210		
	shake right	left			-30	30		
		right			-30	5		
	tilt left	left	4.0	10.0	210	250	300	500
		right			140	170		
	tilt right	left			10	40		
		right			300	330		
3	nod	both	1.5	4.0	70	110	100	200
	shake left	left	3.0	6.0	-5	30	200	350
		right			-30	30		
	shake right	left			150	210		
		right			150	185		
	tilt left	left	4.0	10.0	30	70	400	600
		right			220	350		
	tilt right	left			190	220		
		right			120	150		

Table 1. Ranges defined in the algorithm for the head gesture detection.

head movements could theoretically result in false gesture detection; however, these movements are typically much slower than those required for the head gesture production. However, further research is needed in order to inspect the gesture detection method's resistance to false detections and robustness when users are busy carrying out their natural everyday tasks.

PARTICIPANTS, EQUIPMENT AND PROCEDURE

Eleven volunteers (mean age is 37.3, σ=10.4%, students or staff members at the local university) took part in the test. Six had previous experience in using dwell time for selection in gaze-based interfaces, four were wearing glasses. The tests were organized in the university laboratory and each test session lasted about 15-20 minutes. Tobii T60 eye-tracking device with screen resolution of 1280x1024 pixels was used in this study, and experimental software was developed for gesture detection and stimuli

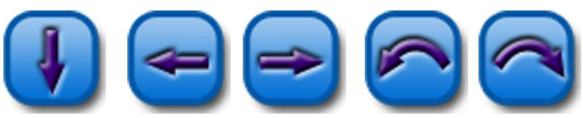

Figure 2. Instructions (targets) displayed during the test.

presentation.

Prior to the test, a supervisor explained the purpose of the study and showed the head gestures to be made. Small amplitude and high speed of the gestures were emphasized when giving the instructions. Then participants had an opportunity to practice the production of the gestures while observing their own eye movements in the camera view, showing feedback for regions to enter by eyes (as illustrated in Figure 1). The training session consisted of 2-4 trials for each gesture.

The eye tracker was calibrated before the test started. During the test, instructions to make a certain gesture were displayed as target icons (50x50 pixels) with an arrow, as shown in Figure 2. Participants were instructed to make only one trial per instruction. The screen was divided into 20 equal cells (5x4), and each target was shown once per cell, and at a random location inside the cell. The order of the targets displayed was randomized. The total number of gestures produced by each participant was 100. The experimental software logged the target location and the required gesture, as well as the average of the gaze points over 200 ms just before the gesture started (GP_{200}, average of the left and the right eye), gaze-on-target time interval, trial duration, and the gesture recognized by the RBA.

After the test, participants were asked about their age, experience in using gaze-controlled interface with dwell-time selection, and fatigue caused by the test. The experienced participants also evaluated the tested gestures for their potential usefulness in HCI with gaze input.

RESULTS AND DISCUSSION

Values of six variables were computed from the collected data: correct detections (TP), absence of detection (FN), misdetections (FP), target hit rate (HIT), an interval between target appearance and gesture recognition (GT), and an interval between target appearance and the first sample landing onto this area (GO). The HIT rate was expressed as the percentage of successful trials when GP_{200} lands onto a 100x100 pixels square with the target at center. The data collected had very little 'gaze-lost' events, and they did not influence the computed values.

The grand means with the standard deviations of TP, FN, FP and HIT for each gesture are shown in Figure 3. The nodding gesture was the best detected (~93.4%, σ=9.5%), and turning and tilting right were the worst (81.8%, σ=18.5% and 71.4%, σ=26.3% respectively). A detailed analysis of the data revealed that some participants tended to overshoot when making other gestures than nodding, and the right eye was often lost by the eye tracker when producing right-directed gestures. Most likely, a better lighting setup would allow equalizing these rates with those observed for left-oriented gestures (~85%, σ =11-14%).

The FP rate was rather low, 1.55% in average (σ=2.64%); the nodding gesture was never interpreted as another gesture. The HIT rate and GO interval were, as expected,

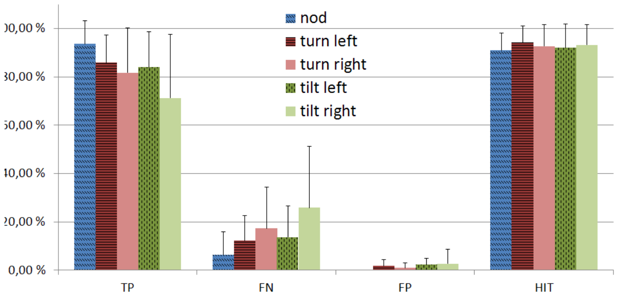

Figure 3. Grand means of TP, FN, FP and HIT for each gesture.

not dependent on the gesture: 92.7% (σ=8.25%) and 329 ms (σ=44 ms) accordingly. The GT interval was about 1 second for nodding, 1.2 seconds for turning, and.1.46 seconds for tilting (σ=120-140 ms). Since nodding took only one second to produce after the target appearance, this gesture would be competitive or even a faster selection method for dwell times of 700 ms or longer (focusing on the target took about 330 ms).

The TP and FN variables were not dependent on the target position. Wrong gesture detection (FP), being a seldom event, occurred usually when a target was located in the lower part of the screen. The HIT rate was about 10% lower than average for the targets displayed on the left screen side. The GT was 50-100 ms shorter for targets located in the central cells compared to targets located in the peripheral cells.

The effect of glasses and the participants' previous experience using gaze-controlled interfaces on TP, FN, FP, HIT and GT was estimated, although the number of participants in each group was rather low (4-7) to make a reliable statistical test. Nevertheless, the test revealed some effect of experience on the HIT rate (94.8% vs. 88.4%, $p=0.051$) and GT (1326 ms vs. 1189 ms, $p=0.05$) when comparing experienced participants against novices. Differences in TP (80.2% vs. 87.2%), FN (17.8% vs. 11.8%) and FP (2% vs. 1%) were recognized as not significant. There was no effect of glasses on any dependent variable; therefore, wearing them did not increase the risk of losing eye(s) by the eye tracker due to head movements.

Participants reported mild fatigue after the test: 2.4 points on the scale from 1 to 7 (2.6 for novices, 2.2 for experienced). The rest of the questionnaire was filled only by the participants experienced in using dwell time as selection method in gaze-controlled interfaces.

We asked the experienced participants to evaluate the potential usefulness of the tested head gestures for selection task for certain applications. Our assumption was that at least one gesture in the test pool would be rated as potentially useful. The five-point Likert scale was used for the evaluation: *impossible* (1), *probably impossible* (2), *hard to say* (3), *probably possible* (4), and *possible* (5). The average rate was 3 (*hard to say*) when asked about using head gestures for completing frequent-selection tasks like typing, 4.3 (*probably possible*) for completing moderate-selection tasks like drawing, and 5 (*possible*) for completing seldom-selection tasks like playing chess. The evaluation shows that the experienced participants positively evaluated head gestures used for interaction in occasional, infrequent-selection tasks. However, the suggestion of using the gestures in frequent-selection tasks was not rejected either.

All experienced participants pointed to the nodding gesture when asked about the best head gesture for making selections. The average rate was 3.5 on scale from 1 to 5 (worse – probably worse – hard to say – probably better – better) when asked to compare the nodding gesture as a selection method against dwell time. Therefore, the subjective evaluation of the nodding gesture is optimistic but cautious, which is obvious as the direct comparison was not performed in this experiment.

When asked whether any of the tested gestures could be used for the navigation tasks or switching between gaze operational modes, four participants pointed to the turning gesture as the best for navigation, and tilting as the best for switching operational modes (other gestures in each case got a single vote).

CONCLUSIONS AND FUTURE WORK

The combined use of gaze for pointing and head gestures as commands has been previously recognized as one of the possibilities for hands-free interaction with computers but it has not been studied much. We conducted a small-scaled study to explore the potential usefulness and usability of this combination when only one video system is needed for such interaction. This is an important issue as both the eye position (used for detection of the head gestures) and the gaze position can be estimated within the same video analysis routine, thus allowing to keep the hardware setup simple and ensure minimal use of computational resources.

The algorithm used for the head gesture detection from eye position in camera view of Tobii T60 eye tracker proved to be usable, although the detection rate of some gestures could be improved. The detection rate of nodding gesture was about 93%, and is close to the best detection rate of the algorithms based on HMM [12]. The eye tracker tended to sometimes lose the gaze of some participants when they were making right-directed gestures, but we expect that better set lightning conditions should help to avoid this problem.

One hundred gestures within 4-5 minutes caused only a little fatigue for participants. The participants who were experienced in using dwell time in gaze controlled interfaces were quite positive about the proposal to replace the dwell time by the nodding gesture, especially occasional infrequent tasks. The nodding gesture took only 1 second to produce (of which focusing on target took about 330 ms), therefore, using it instead of dwell time longer than 700 ms may result in faster and more convenient interaction.

A direct comparison between dwell time and nodding gesture as selection methods is the obvious continuation of this study. The gesture detection algorithm could be improved by analyzing the ratio between the amplitude and time interval of eye movement: the patterns of this ratio for a single gesture may have smaller variation between trials and subjects.

Although not all people with physical limitations may be able to use the proposed method, the combined use of gaze pointing and head gestures could be highly useful in multimodal applications. We believe it to be especially useful in transient interactions in pervasive and mobile environments where the movement of the head and the body is part of the natural interaction. Further research is required to study its applicability in different applications and varying contexts.

ACKNOWLEDGMENTS
We wish to thank Poika Isokoski and other members of the TAUCHI Visual Interaction Research Group for their invaluable feedback on this research.

REFERENCES

1. Bee, N. and André, E. (2008). Writing with Your Eye: A Dwell Time Free Writing System Adapted to the Nature of Human Eye Gaze. In *Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems (PIT '08),* Springer-Verlag, Berlin, Heidelberg, 111-122.
2. Donegan, M., Morris, J.D., Corno, F., Signorile, I., Chió, A., Pasian, V., Vignola, A., Buchholz, M., and Holmqvist, E. (2009). Understanding users and their needs. *Universal Access in Information Society 8(4)*, 259-275.
3. Drewes, H. and Schmidt, A. (2007). Interacting with the computer using gaze gestures. In *Proceedings of the 11th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT'07),* Vol. 2, Springer-Verlag, Berlin, Heidelberg, 475-488.
4. Gizatdinova, Y., Špakov, O., and Surakka, V. (2012). Face typing: Visual gesture-based perceptual interface for typing with a scrollable virtual keyboard. In *Proceedings of the 2012 IEEE Workshop on the Applications of Computer Vision (WACV'12)*, IEEE Computer Society, 7 pages (in press).
5. Hansen, D. W., Skovsgaard, H. H. T., Hansen, J. P., and Møllenbach, E. (2008). Noise tolerant selection by gaze-controlled pan and zoom in 3D. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications (ETRA'08)*, ACM, New York, NY, USA, 205-212.
6. Heikkilä, H., and Räihä, K.-J. (2009). Speed and Accuracy of Gaze Gestures. *Journal of Eye Movement Research*, 3(2):1, 1-14.
7. Hornof, A. J., and Cavender, A. (2005). EyeDraw: enabling children with severe motor impairments to draw with their eyes. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'05)*, ACM, New York, NY, USA, 161-170.
8. Huckauf, A. and Urbina, M. H. (2008). On object selection in gaze controlled environments. *Journal of Eye Movement Research*, 2(4):4, 1-7.
9. Huckauf, A. and Urbina, M. H. (2011). Object selection in gaze controlled systems: What you don't look at is what you get. *ACM Transactions on Applied Perception*, 8(2):13, 14 pages.
10. Isokoski, P. (2000). Text input methods for eye trackers using off-screen targets. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA'00)*, ACM, New York, NY, USA, 15-21.
11. Istance, H., Bates, R., Hyrskykari, A., and Vickers, S. (2008). Snap clutch, a moded approach to solving the Midas touch problem. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications (ETRA'08)*, ACM, New York, NY, 221–228.
12. Kang, Y.G., Joo, H.J., and Rhee, P. (2006). Real Time Head Nod and Shake Detection Using HMMs. In *Proceedings of Knowledge-Based Intelligent Information and Engineering Systems*, Vol. 4253, Springer, 707-714.
13. Kumar, M., Paepcke, A., and Winograd, T. (2007). EyePoint: practical pointing and selection using gaze and keyboard. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'07)*, ACM, New York, NY, USA, 421-430.
14. Lin, C. S., Ho, C. W., Chan, C. N., Chau, C. R., Wu, Y. C., and Yeh, M. S. (2007). An eye-tracking and head-control system using movement increment-coordinate method. *Journal of Optics & Laser Technology*, 39(6), 1218-1225.
15. Majaranta, P., Ahola, U.-K., and Špakov, O. (2009). Fast Gaze Typing with an Adjustable Dwell Time. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'09)*, ACM, New York, NY, USA, 357-360.
16. Majaranta, P., and Räihä, K-J. (2002). Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 symposium on Eye tracking research & applications (ETRA'02)*, ACM, New York, NY, USA, 15-22.
17. Mardanbegi, D., Hansen, D. W. and Pederson, T. (2012). Eye-based head gestures. In *Proceedings of the 2012 Symposium on Eye Tracking Research and Applications (ETRA '12)*, ACM, New York, NY, USA, 139-146.
18. Møllenbach, E., Lillholm, M., Gail, A., and Hansen. J. P. (2010). Single gaze gestures. In *Proceedings of the 2010 Symposium on Eye-Tracking Research &*

Applications (ETRA'10), ACM, New York, NY, USA, 177-180.

19. Nonaka, H. (2003). Communication Interface with Eye-Gaze and Head Gesture Using Successive DP Matching and Fuzzy Inference. *Journal of Intelligent Information Systems*, 21(2), Kluwer Academic Publishers Hingham, MA, USA, 105-112.
20. Rasmusson, D., Chappell, R., and Trego, M. (1999). Quick Glance: Eye-tracking Access to the Windows95 Operating Environment. In *Proceedings of the Fourteenth International Conference on Technology and Persons with Disabilities (CSUN'99)*, Los Angeles, CA.
21. Skovsgaard, H., Räihä, K.-J., and Tall, M. (2012). Computer Control by Gaze. In Majaranta et al. (Eds.), *Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies* (pp. 63-77), IGI Global.
22. Špakov, O., and Miniotas, D. (2005). EyeChess: A Tutorial for Endgames with Gaze-Controlled Pieces. In *Proceedings of Communication by Gaze Interaction (COGAIN'05)*, 16-18.
23. Surakka, V., Illi, M., and Isokoski, P. (2004). Gazing and frowning as a new human--computer interaction technique. *ACM Transactions on Applied Perception*, 1(1), 40-56.
24. Tan, W., and Rong, G. (2003). A real-time head nod and shake detector using HMMs. *Expert Systems with Applications*, 25, 461-466.
25. Tuisku, O., Majaranta, P., Isokoski, P., and Räihä, K-J. (2008). Now Dasher! Dash away!: longitudinal study of fast text entry by Eye Gaze. *In Proceedings of the 2008 Symposium on Eye Tracking Research & Applications (ETRA'08)*, ACM, New York, NY, USA, 19-26.
26. Velichkovsky, B., Sprenger, A. and Unema, P. (1997) Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem". In *Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 1997)*, 509-516. London: Chapman and Hall
27. Ward, D. J. and MacKay, D. J. C. (2002) Fast hands-free writing by gaze direction. Nature 418(6900), 838.
28. Wobbrock, J. O., Rubinstein, J., Sawyer, M. W., and Duchowski, A. T. (2008). Longitudinal evaluation of discrete consecutive gaze gestures for text entry. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications (ETRA'08)*, ACM, New York, NY, USA, 11–18