2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Multimodal Direct Gaze Interface for Wheelchairs and Teleoperated Robots

Isamu Poy[1*], Liang Wu[2*], and Bertram E. Shi[2], *Fellow, IEEE*

Abstract— **Gaze-based interfaces are especially useful for people with disabilities involving the upper limbs or hands. Typically, users select from a number of options (e.g. letters or commands) displayed on a screen by gazing at the desired option. However, in some applications, e.g. gaze-based driving, it may be dangerous to direct gaze away from the environment towards a separate display. In addition, a purely gaze based interface can present a high cognitive load to users, as gaze is not normally used for selection and/or control, but rather for other purposes, such as information gathering. To address these issues, this paper presents a cost-effective multi-modal system for gaze based driving which combines appearance-based gaze estimates derived from webcam images with push button inputs that trigger command execution. This system uses an intuitive ”direct interface”, where users determine the direction of motion by gazing in the corresponding direction in the environment. We have implemented the system for both wheelchair control and robotic teleoperation. The use of our system should provide substantial benefits for patients with severe motor disabilities, such as ALS, by providing them with a more natural and affordable method of wheelchair control. We compare the performance of our system to the more conventional and common ”indirect” system where gaze is used to select commands from a separate display, showing that our system enables faster and more efficient navigation.**

I. Introduction

Eye tracking has been used successfully as an input modality for human-robot [1], [2], [3], and human-computer [4], [5], [6], [7], [8] interfaces. In particular, gaze-based interfaces have been proposed for driving robotic wheelchairs and tele-operated robots. Users view the environment around the vehicle, decide how to control it, and then issue commands to execute a desired trajectory. In wheelchair control, the user is seated in the vehicle and views the environment directly. In tele-operation, the user is seated remotely, and views the environment through a video feed.

Much attention has been placed on pure gaze-based interfaces. These have the advantage that they do not rely on any other cues or actions by the user, imposing minimal constraints and prior assumptions. On the other hand, relying only upon gaze for control leads to awkward interfaces and interactions, because the most common use of gaze is for gathering information about the environment, not control. The most common problem arising from this mismatch is the Midas-Touch problem, i.e. unintentional selection, in gaze-based interfaces for selection. The most common way to deal with this is by a fixed dwell time, during which users must fixate on the desired option.

The easiest way to eliminate dwell time is to confirm selection of the fixated object using a different modality such as EEG [9], EMG [10], or touch [11]. However, these approaches impose additional assumptions that reduce the size of the user base. They may have also have additional drawbacks, such as noise [12]. Multi-modality expands the design space, but effective design remains a challenging problem.

This paper describes a system for gaze-based driving where users indicate the direction they wish to move by gazing in that direction, and trigger motion using a switch. We estimate eye gaze from a remote webcam using an appearance based gaze estimator. We feel that this is a particularly appealing combination. Specifying commands directly in the environment reduces gaze shifts, enabling the user to remain fully engaged in the navigation task. Previous direct methods for low level control [13][14] were based on continuous control. By relying on an additional switch, our method enables users to monitor the environment during navigation. This is less taxing and more consistent with normal gaze during navigation. Previous work with multimodal interfaces, e.g. using Brain Computer Interfaces (BCI) [9] or buttons [11], did not use environmentally centered gaze. We compare our interface with a multimodal indirect interface similar to that proposed by Meena et al. [11]. Our system results in faster and more efficient navigation.

II. Related Work

To clarify our system’s uniqueness, Table I classifies it and past work along multiple dimensions.

The first dimension is the interface location, i.e. where do users look to indicate intent. In direct interfaces, users look directly into the environment towards a desired location, object or direction of motion. In indirect interfaces, users look at a display screen, usually located somewhere in the periphery to avoid blocking the view of the environment. While they enable selection from more commands, they require frequent gaze shifts away from the environment. Overlay interfaces are a new middle ground enabled by VR/AR headsets. These interfaces overlay a command display over the environmental view, reducing gaze shifts away from the environment. However, the headset may

[*]indicates equal contribution

[1]IP is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada. Please direct inquiries to isamu.poy@mail.utoronto.ca.

[2]LW and BES are with Department of Electronic and Computer Engineering, the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong SAR. This work was supported in part by the Hong Kong Innovation and Technology Support Programme under grant ITS/406/16FP.

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

Authors	Interface Loc.			Timing of Cmd				Cmd Level		Detect Method			Track. Loc.	
	dir	ind	over	cont	dwell	mult	gest	high	low	act	app	EOG	rem	head
Al-Haddad et al.[15]	✓						✓		✓			✓		✓
Cruz et al.[9]		✓				✓			✓			✓		✓
Araujo et al. cont.[16]			✓	✓					✓	✓				✓
Araujo et al. (dwell) [16]			✓		✓			✓		✓				✓
Araujo et al. (waypoint) [16]			✓	✓				✓		✓				✓
Eid et al.[17]		✓					✓		✓	✓			✓	
Meena et al.[11]		✓				✓			✓		✓		✓	
Singer et al.[18]			✓		✓				✓	✓	✓			
Tall et al.[13]	✓			✓					✓		✓		✓	
Yuan et al.[2]	✓				✓			✓		✓				✓
Zhang et al.[14]	✓			✓					✓	✓				✓
Proposed Direct	✓					✓			✓		✓		✓	

TABLE I: Comparison of gaze-based control interfaces based on several dimensions: where the interface is located, the initiation of gaze control, command level, gaze estimation method, and eye tracker location. Red text indicates the systems presented in this paper. The table uses the following abbreviations: dir = direct interface, ind = indirect interface, over = overlay interface, cont = continuous gaze, dwell = dwell based, mult = multimodal interface, gest = gesture based, low = manual directions (L,R,F,B), act = gaze estimation device, app = appearance based gaze, rem = remote (webcam), head = head mounted, Loc = Location, Cmd = Command, Track = Tracker

be expensive and uncomfortable. Singer et al. proposed a headset-free alternative, which used an acrylic frame placed in front of the user [18].

The second dimension describes how commands are initiated. In a continuous interface, commands are issued immediately based on the instantaneous gaze location. Stop commands are often issued by having user close his/her eyes or by looking away. While responsive, they can be fatiguing, as they do not allow the user to scan the environment during control. Dwell based interfaces select commands after the users fixate them for a fixed amount of time. While these avoid the Midas Touch problem, the delay introduced means they should be avoided for emergency actions like braking. Gesture-based initiation relies upon eye cues other than gaze direction, e.g. blinking, to initiate commands. Multimodal interfaces rely upon cues other than eye-gaze to initiate commands. For example, users with motor impairment can often still indicate binary intent (start/stop or yes/no) using switches controlled by their hand, foot, mouth, head, etc.

The third dimension is the level of the command. A high-level command is typically a specific location in the environment where the user wants to move to. These rely upon autonomous navigation technology, such as simultaneous localization and mapping and route planning and execution, in order to complete the commands. Low level commands correspond to specific motions of the vehicle. These may be either discrete (e.g. forward, left, right, back, stop) commands or continuous (e.g. rotational or linear speeds). Low-level interfaces are more involved, but are useful for precise navigation.

The fourth dimension is the method used to estimate eye gaze. Options here include active illumination (e.g. pupil centre corneal reflection or PCCR) eye trackers, which are accurate but more expensive as they require specialized hardware; appearance-based eye trackers, which are less accurate but lower cost as they rely only upon using commonly available webcams; and electrooculography (EOG)-based eye trackers, which are the most invasive as they require electrodes attached to the skin. Although PCCR eye trackers from companies like Tobii, are becoming lower in cost, they are still not as ubiquitous and low cost as webcams.

The final dimension is the location of the eye tracker, i.e. whether it is located in the environment remote from the user, or whether it is attached to the user, usually the head, but sometimes the body. Remote eye trackers are less intrusive, but constrain user movement more.

As shown in Table I, our proposed interface extends previous work by exploring a new point in the design space, which brings the unique advantages described previously.

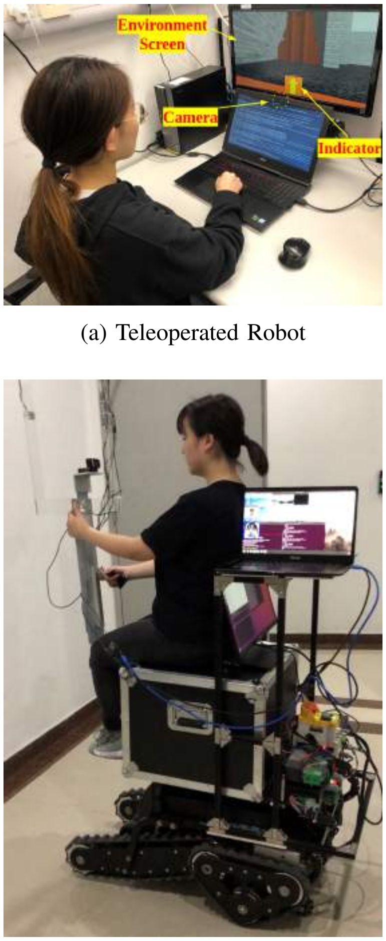
III. Methodology

A. Multimodal Gaze Interface

We have used our proposed direct gaze interface to control the robotic platforms shown in Fig. 1.

Fig. 1a shows the robotic teleoperation platform. The large monitor shows the view of the environment in front of the robot. Feedback about the direction of motion indicated by the instantaneous gaze location is provided using an arrow located at the bottom of the screen. The user can monitor this feedback using their peripheral vision while maintaining their gaze inside the environment. This enables the user to explore and monitor the environment continuously as they navigate the vehicle. The camera on the laptop below the large screen acquires images of the user's face for input to the gaze estimator. The screen of the laptop is not used for the direct gaze interface. However, it is used in the indirect gaze interface described below.

Fig. 1b shows the wheelchair platform. The user can view the environment directly. Images of the user's face are acquired using the black webcam located in front of and slightly below eye level of the user, so that it does not



(a) Teleoperated Robot

(b) Wheelchair

Fig. 1: Robotic platforms controlled by gaze interface

block the user's view. Feedback about the direction of motion indicated by gaze are provided by light emitting diodes mounted on a transparent acrylic frame in front of the user. The user does not need to fixate on the feedback indicators, but rather can monitor them using peripheral vision.

Subjects select from one of four directions to move (forward, left, right, and backwards) by gazing in the appropriate direction in the environment. The visual field in front of the user is divided into four regions, one for each of the commands. This allows the user to gaze freely in the general direction they would naturally look during navigation using other modalities, such as a joystick or keyboard. For example, when rotating the vehicle to the right, users will naturally look somewhere to the right.

For example, in the teleoperation system, the forward and backwards commands were indicated when the user gazed at the top and bottom sections of the screen respectively. The selection box dimensions were half the screen width and a quarter of the screen height. Similarly, the rotate left and rotate right command selection boxes were a quarter of the screen width and the full screen height. Our initial informal experiments with the gaze interface suggested that this segmentation was the most intuitive.

Gaze was only used to indicate the intended command. Command execution was triggered by the user's press of a button, which sent the intended command indicated by the instantaneous gaze at the time of the button press to the robot and maintained that command for the duration of the button press irrespective of the user's gaze after the command initiation. Releasing the button caused the robot to stop.

In our experiments, the button was a key on the keyboard, but it could be implemented in any number of ways (e.g. foot switch, puff switch) depending on the residual function of user. While the additional input adds some restriction, this is minimized by the flexibility of its implementation.

Gaze estimates used for selection are derived from images acquired from webcams placed so as to not block the user's normal view. The remote webcam is less intrusive than overlay technologies which rely upon virtual or augmented reality. It is also lower in cost than eye trackers requiring active illumination, and more applicable in outdoor environments, where sunlight can overwhelm the infrared light used by PCCR eye trackers.

As benchmarks for comparison, we also implemented two other gaze interfaces in our teleoperation system: a indirect gaze-based system and a keyboard only system. In the indirect gaze-based teleoperation system, users selected from four direction commands displayed in labelled squares on the laptop screen below the large monitor in Fig. 1a. Gazing at a command caused the corresponding square to change color, but did not trigger execution. As in the direct interface, command execution was triggered and maintain by a key press. In the keyboard only system, the four commands were mapped to four keys on the keyboard, and were executed immediately once the corresponding key was pressed.

B. Experimental Setup and Task Description

In our experiments, we used the teleoperation interface to control a robot simulated using the Robot Operating System (ROS) in two Gazebo environments. This enabled us to collect more detailed synchronized information about gaze behavior and the robot trajectory than possible using the wheelchair set-up. The simulations were run on a desktop computer equipped with a 4GB NVIDIA GTX Graphics Processing Unit (GPU), which ran the image processing and gaze-estimation algorithms built using the Tensorflow and NVIDIA CUDnn libraries. We used a gaze estimator developed previously by our lab [19].

Fig. 2 shows how the interface interacts with the simulation. Images of the user captured by the web camera are passed to the appearance based gaze estimator. Gaze estimates are generated in camera coordinates, then transformed to screen coordinates using a transformation matrix estimated in a prior calibration stage. For the gaze interfaces, commands were indicated by the gaze location in the view of the environment (for the direct interface) or on the separate command interface screen (for the indirect interface), and initiated/maintained by the button press. Commands were sent to the ROS simulator, which used the

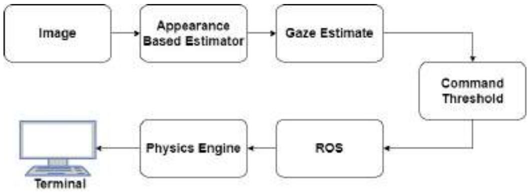
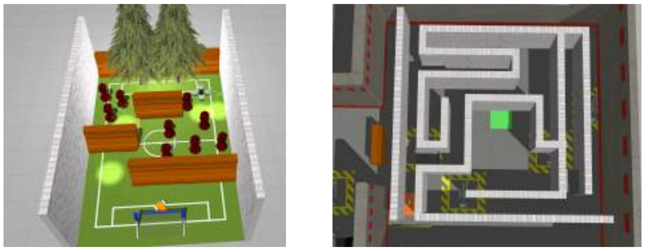


Fig. 2: System Diagram



(a) Obstacle Environment

(b) Maze Environment

Fig. 3: Simulated Environments

built-in physics engine to update the robot position and to render views of the 3D environment. When the key was released, the stop command was sent to the ROS simulator.

Seven subjects participated in this experiment. All were university students with prior experience using eye trackers. At the beginning of the experiment, the experimenter explained to the subject the operation of each interface.

Each subject performed six navigation trials: one for each combination of the three interfaces in the two environments. In each trial, the subject was asked to drive the robot from start to a destination as quickly as possible. Figure 3 shows the two environments. In these figures, the robot is in the starting position and the green square indicates the destination. The order of the interface and environments were randomized across users. After the subject reached the goal, they were allowed to rest until they were ready for the next trial.

IV. Experimental Results

A. Task Completion Time

Table II shows the average completion times for each interface in the Obstacle and Maze environments. Using the direct interface, subjects were able to navigate 26% and 23% faster in the obstacle and maze environments respectively, than when using the indirect interface. We normalized completion times by the time taken by the keyboard interface because normalized values allowed for easier comparison between different interfaces across different environments, where completion times differed. The times from the keyboard interface are a target lower bound for the gaze based interface, as it is the most frequently used interface for controlling characters in video games by non-disabled people.

Fig. 4 plots an example of distance travelled versus time using the three interfaces by a subject in one of the environments. Steeper curves correspond to faster movement.

Obstacle Environment		
Category	Time (s)	Normalized
Direct	181.0	123%
Indirect	218.6	149%
Keyboard	146.3	100%

Maze Environment		
Category	Time (s)	Normalized
Direct	467.7	145%
Indirect	540.4	168%
Keyboard	321.2	100%

TABLE II: Normalized average speed results

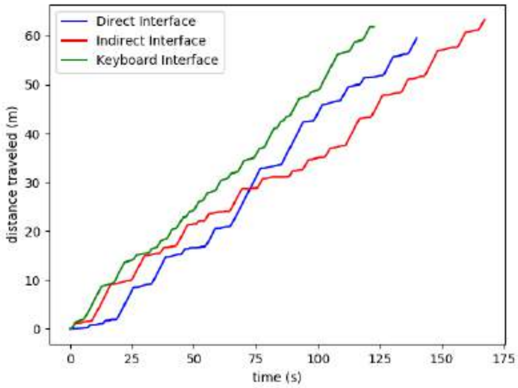


Fig. 4: Distance vs. Time graph of different interfaces

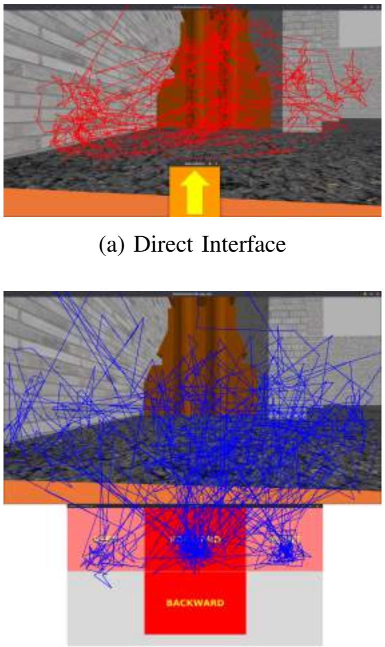
With the direct interface, the subject performed fewer turns than with the indirect interface. This suggests that the subject was able to navigate more efficiently when s/he could keep his/her gaze in the environment, than when switching back and forth between the environment and a separate display.

Fig. 5 shows gaze trajectories generated by an example subject over an example trial for the two interfaces. For the direct interface (Fig. 5a) gaze points are mostly concentrated in the environment. For the indirect interface (Fig. 5b), gaze frequently switched between the control panel and the environment. We hypothesize that these frequent switches added additional overhead, as the subject needed to re-orient his/her gaze in the environment after each gaze switch, making navigation more difficult. This suggests that when designing a gaze-based human-robot interaction system, developers should seek to avoid gaze switching.

Table III shows the results of a Two-Way ANOVA with replication of the speed data. The two factors were ”Environment” and ”Interface”. There was a significant effect of both interface and environment ($p < 0.03$). We observed no interaction between interface and environment, indicating that the performance gain by the direct interface does not depend upon the environment.

V. Conclusion

This paper demonstrated a low cost gaze interface which can be applied to both robotic wheelchairs and teleoperated robots. Our experimental results show that our direct environmentally grounded interface enables more efficient and faster navigation than an indirect interface, which is



(a) Direct Interface

(b) Indirect Interface

Fig. 5: Example gaze traces

ANOVA						
Variation	SS	df	MS	F	P-value	F crit
Environment	0.37	1	0.37	22.18	<0.03	4.171
Interface	0.68	2	0.34	20.31	<0.03	3.316
Interaction	0.08	2	0.04	2.390	0.109	3.316
Within	0.50	30	0.02			
Total	1.64	35				

TABLE III: Two-way ANOVA w/ Replication on avg. speeds

more commonly used. The direct interface allows the user to maintain gaze in the environment during driving, rather than gazing away to select commands.

Our work can be extended in several ways. First, we can incorporate gaze prediction in addition to gaze estimation, as it has been suggested that its addition can enable more accurate commands [20], [21]. Second, our system can be tested on disabled subjects in wheelchairs. It is possible that patients will take longer to navigate than healthy subjects [22], but we expect that the relative advantages of the direct versus indirect interface will be maintained.

References

[1] A. Frisoli, C. Loconsole, D. Leonardis, F. Banno, M. Barsotti, C. Chisari, and M. Bergamasco, “A new gaze-bci-driven control of an upper limb exoskeleton for rehabilitation in real-world tasks,” in *IEEE Transactions on Systems, Man, and Cybernetics, Part C*. New York, NY: IEEE Press, December, 2012 2012, p. 10.

[2] L. Yuan, C. Reardon, G. Warnell, and G. Loianno, “Human gaze-driven spatial tasking of an autonomous mav,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1343 – 1350, April 2019.

[3] H. O. Latif, N. Sherkat, and A. Lotfi, “Teleoperation through eye gaze (telegaze): A multimodal approach,” in *International Conference on Robotics and Biomimetics*. New York, NY: IEEE Press, Dec, 2009 2009.

[4] I. F. Ince and J. W. Kim, “A 2d eye gaze estimation system with low-resolution webcam images,” in *EURASIP Journal on Advances in Signal Processing*. New York, NY: Springer, August, 2011 2011, p. 11.

[5] Y.-T. Lin, R.-Y. Lin, Y.-C. Lin, and G. C. Lee, “Real-time eye-gaze estimation using a low-resolution webcam,” in *Multimedia Tools and Applications*. New York, NY: Springer, August, 2011 2012, p. 11.

[6] J. Pi and B. E. Shi, “Task-embedded online eye-tracker calibration for improving robustness to head motion,” in *ETRA ’19*. New York, NY: ACM Press, June, 2019 2019, p. 9.

[7] Z. Chen and B. E. Shi, “Using variable dwell time to accelerate gaze-based web browsing with two-step selection,” *International Journal of Human-Computer Interaction*, vol. 35, no. 3, pp. 240 – 255, March 2019.

[8] S. Vickers, H. Istance, and A. Hyrskykari, “Performing locomotion tasks in immersive computer games with an adapted eye-tracking interface,” in *ACM Transactions on Accessible Computing*. New York, NY: ACM Press, Sept, 2013 2013.

[9] R. Cruz, V. Souza, T. B. Filho, and V. L. Jr., “Electric powered wheelchair command by information fusion from eye tracking and bci,” in *IEEE International Conference on Consumer Electronics (ICCE)*. New York, NY: IEEE Press, Jan, 2019 2019, p. 2.

[10] J. C. Mateo, J. S. Agustin, and J. P. Hansen, “Gaze beats mouse: Hands-free selection by combining gaze and emg,” in *CHI 2008 Proceedings - Works In Progress*. New York, NY: ACM Press, April 5-10 2008, p. 6.

[11] Y. K. Meena, H. Cecotti, K. Wong-Lin, and G. Prasad, “A multimodal interface to resolve the midas-touch problem in gaze controlled wheelchair,” in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. New York, NY: IEEE Press, July 11-15, 2008 2017, p. 4.

[12] K. S. Türker, “Electromyography: Some methodological problems and issues,” *Physical Therapy*, vol. 73, no. 10, pp. 698 – 710, October 1993.

[13] M. Tall, J. P. Hansen, A. Alapetite, D. W. Hansen, J. S. Agustin, E. Mollenbach, and H. H. Skovsgaard, “Gaze-controlled driving,” in *CHI 2009 Spotlight on Works in Progress Session 2*. New York, NY: ACM Press, April 4-9, 2009 2009, p. 5.

[14] G. Zhang and J. P. Hansen, “A virtual reality simulator for training gaze control of wheeled tele-robots,” in *VRST’19*. New York, NY: ACM Press, November 12-15, 2019 2019, p. 2.

[15] A. Al-Haddad, R. Sudirman, and C. Omar, “Gaze at desired destination, and wheelchair will navigate towards it. new technique to guide wheelchair motion based on eog signals,” in *International Conference on Informatics and Computational Intelligence (ICI)*. New York, NY: IEEE Press, Dec, 2011 2011.

[16] J. de Araujo, J. P. Hansen, G. Zhang, and S. Puthusserypady, “Exploring eye-gaze wheelchair control,” in *ETRA’20 Adjunct*. New York, NY: ACM Press, June 2–5, 2020 2020, p. 2.

[17] M. A. Eid, N. Giakoumidis, and A. E. Saddik, “A novel eye-gaze-controlled wheelchair system for navigating unknown environments: Case study with a person with als,” *IEEE Access*, vol. 4, no. 5, pp. 558–573, Nov. 2016.

[18] C. Singer and B. Hartmann, “See-thru: Towards minimally obstructive eye-controlled wheelchair interfaces,” in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. New York, NY: ACM Press, October, 2019 2019, p. 10.

[19] Z. Chen and B. E. Shi, “Appearance-based gaze estimation using dilated-convolutions,” in *Asian Conference on Computer Vision 2018*. New York, NY: Springer, May 26, 2019 2019, p. 15.

[20] P. Novák, T. Krajník, L. Přeučil, M. Fejtová, and O. Štěpánková, “Ai support for a gaze controlled wheelchair,” in *The 4th Conference on Communication by Gaze Interaction – COGAIN 2008*. Seattle, WA: Semantic Scholar, September 2-3, 2008 2008, p. 4.

[21] R. Bednarik, H. Vrzakova, and M. Hradis, “What do you want to do next: A novel approach for intent prediction in gaze-based interaction,” in *ETRA ’12*. New York, NY: ACM Press, March, 2012 2012, p. 8.

[22] K. Vaccaro, D. Huang, M. Eslami, C. Sandvig, K. Hamilton, and K. Karahalios, “The illusion of control: Placebo effects of control settings,” in *CHI ’18*. New York, NY: ACM Press, April 2012 2018, p. 13.