Enhancing Mobile Voice Assistants with *WorldGaze*

Sven Mayer[1] **Gierad Laput**[2,1] **Chris Harrison**[1]
[1]Human-Computer Interaction Institute
Carnegie Mellon University, Pittsburgh, PA, USA
{sven.mayer, chris.harrison}@cs.cmu.edu
[2]Apple Inc.
One Apple Park Way, Cupertino, CA, USA
gierad@apple.com

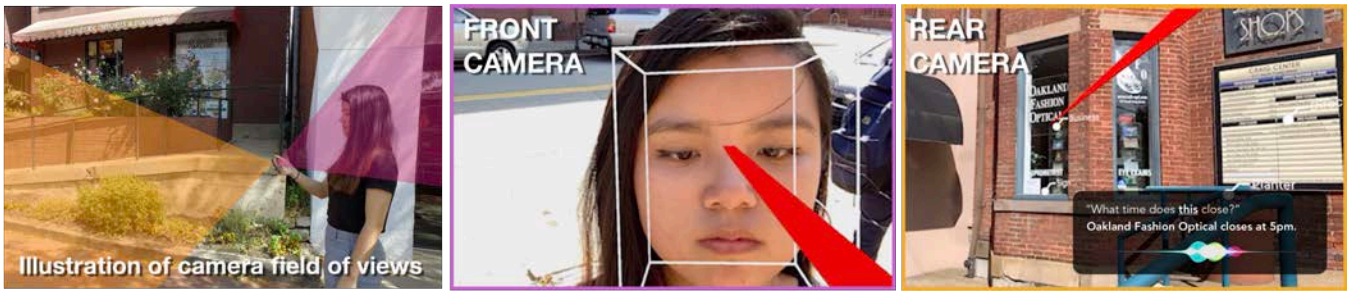

Figure 1. WorldGaze simultaneously opens the front and rear camera on smartphones. The front camera is used to track the user's 3D head vector, which is then raycast into the world as seen by the rear camera. This allows users to intuitively define an object or region of interest using their head gaze, which voice assistants can utilize for more precise and natural interactions (right bottom).

ABSTRACT
Contemporary voice assistants require that objects of interest be specified in spoken commands. Of course, users are often looking directly at the object or place of interest – fine-grained, contextual information that is currently unused. We present WorldGaze, a software-only method for smartphones that provides the real-world gaze location of a user that voice agents can utilize for rapid, natural, and precise interactions. We achieve this by simultaneously opening the front and rear cameras of a smartphone. The front-facing camera is used to track the head in 3D, including estimating its direction vector. As the geometry of the front and back cameras are fixed and known, we can raycast the head vector into the 3D world scene as captured by the rear-facing camera. This allows the user to intuitively define an object or region of interest using their head gaze. We started our investigations with a qualitative exploration of competing methods, before developing a functional, real-time implementation. We conclude with an evaluation that shows WorldGaze can be quick and accurate, opening new multimodal gaze+voice interactions for mobile voice agents.

Author Keywords
WorldGaze; interaction techniques; mobile interaction.

CSS Concepts
• **Human-centered computing~Interaction techniques**

CHI '20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00
https://doi.org/10.1145/3313831.3376479

INTRODUCTION
Today's voice assistants lack fine-grained contextual awareness, requiring users to be unambiguous in their voice commands. In a smart home setting, one cannot simply say "turn that up" without providing extra context, even when the object of use would be obvious to humans in the room (*e.g.,* when watching TV, cooking on a stove, listening to music on a sound system). This problem is particularly acute in mobile voice interactions, where users are on the go and the physical context is constantly changing. Even with GPS, mobile voice agents cannot resolve questions like "when does *this* close?" or "what is the rating of *this* restaurant?" (see Figure 1).

Users are often directly looking at the objects they are inquiring about. This real-world gaze location is an obvious source of fine-grained, contextual information that could both resolve ambiguities in spoken commands and enable more rapid and human-like interactions [51]. Indeed, multimodal gaze+voice input has long been recognized as a potent combination, starting with seminal work in 1980 by Bolt [10]. However, prior gaze-augmented voice agents generally require environments to be pre-registered or otherwise constrained, and most often employ head-worn or fixed sensing infrastructure to capture gaze. This precludes true mobility, especially for popular form factors such as smartphones.

In this work, we aimed to develop a *practical* implementation of ad-hoc, real-world gaze location sensing for use with mobile voice agents. Critically, our implementation is software only, requiring no new hardware or modification of the environment. It works in both static indoor scenes as well as outdoor streetscapes while walking. This is achieved by simultaneously opening the front and rear cameras of a smartphone, offering a combined field of view of just over 200 degrees on the latest generation of smartphones. The front-facing camera is used to track the head in 3D, including

its direction vector (*i.e.,* 6DOF). As the geometry of the front and back cameras is fixed and known, along with all of the lens intrinsics, we can raycast the head vector into the 3D world scene as captured by the rear-facing camera.

This allows the user to intuitively define an object or region of interest using their head gaze. Voice assistants can then use this extra contextual information to make inquiries that are more precise and natural. In addition to streetscape questions, such as "is this restaurant good?", WorldGaze can also facilitate rapid interactions in density instrumented smart environments, including automatically resolving otherwise ambiguous actions, such as "go", "play" and "stop." We also believe WorldGaze could help to socialize mobile AR experiences, currently typified by people walking down the street looking down at their devices. We believe our approach can help people better engage with the world and the people around them, while still offering powerful digital interactions through voice.

We started our investigations with a qualitative study that helped to ground our design and assess user acceptability. With encouraging results, we then moved to development of a functional, real-time prototype, constraining ourselves to hardware found on commodity smartphones. We conclude the paper with a performance evaluation that shows WorldGaze can be quick and accurate, highlighting the potential of multimodal mobile interactions.

RELATED WORK

Tracking a user's gaze for interactive purposes has been the subject of research for many decades. We first position this paper with respect to the large multimodal interaction literature. We then briefly review the gaze tracking literature, focusing on mobile systems, followed by discussion on existing mobile approaches for inferring geospatial and physical context. Finally, we conclude with systems that combine both gaze and voice, which is most relevant to WorldGaze.

Multimodal Interaction

A wide variety of multimodal interaction techniques have been considered that combine two or more input modalities to enable more accurate or expressive interactions. For example, combining pen and finger input on touchscreens has been an area of particular interest, *e.g.,* Cami *et al.* [12], and Hinckley *et al.* [24]. Researchers have also looked at combining touch and gaze for enhanced selection or enabling new functionality, *e.g.,* Pfeuffer *et al.* [36]. Speech combined with gestures [11][37] or handwriting [2] has been used to overcome individual drawbacks. Clark *et al.* [14] offer a comprehensive survey on multimodal techniques incorporating speech input. WorldGaze contributes to this literature, and like most other multimodal techniques, it offers unique interactions that move beyond what speech or gaze can offer alone.

Gaze Pointing

Bolt's pioneering work ("put that there" [10]) used mid-air pointing to select objects at a distance. This paved the way for follow-up research which explored the usability of multimodal input, including deictic hand gestures [47], mid-air pointing [31] [32], and eye tracking, such as MAGIC pointing [50]. Zhai *et al.* [50] took advantage of clutching mechanisms (*e.g.,* mouse + gaze or hand gesture + gaze) for target selection, helping to mitigate the "Midas touch" effect inherent in gaze-driven interactions [24].

Following Bolt and Zhai's seminal work, more sophisticated approaches for gaze pointing have emerged. For instance, Drewes *et al.* [15] proposed using a stationary eye tracker to support mobile phone interactions. Mardanbegi and Hansen [30] extend this idea, using gaze-based selection for wall displays. More recently, Orbits [16] explored a gaze tracking technique based on smooth pursuit coupling, while Schweigert *et al.* [44] investigated gaze direction for targeting and mid-air pointing as a selection trigger. This prior work illustrates the value and feasibility of gaze as an input channel, all of which inspired the direction of our work.

Geospatial Mobile Interactions

Knowledge about a user's physical context is especially valuable for computers that are mobile. Armed with such information, these devices provide users with more timely and contextually relevant information. Technologies like GPS and WiFi localization offer coarse location information that could identify *e.g.,* which Starbucks the user is standing in front of (*i.e.,* city block scale), but they are not precise enough to resolve *e.g.,* which exact business the user is inquiring about without specifying a name.

Bluetooth beacons and ultrasonic localization systems are more targeted, offering room-scale accuracy (or better), which is sufficient to resolve questions with a single applicable local target, such as "what is this printer's name?". However, these techniques fail when there are multiple applicable targets ("turn on"), even when a category is provided ("what model car is this?" when standing in a parking lot). As noted by Schmidt *et al.* [42] "there is more to context than location". We agree and believe gaze to be among the strongest and natural complementary channels.

Object Context + Voice Interactions

Glenn *et al* [20] presented one of the earliest systems combining gaze and voice. Follow-up work has focused on specific tasks, for example, Koons *et al.* [27] built a system combing speech, gaze, and hand gestures for map manipulation, while Acartürk *et al.* [1] proposed using gaze and voice to enable interaction with computers for the elderly. Other researchers have explored using voice within context-specific situations. For example, Roider *et al.* [40] and Neßelrath *et al.* [34] used gaze and eye tracking on dashboards to enable expressive car-based interactions. Regardless of the context, researchers have shown that multi-modal systems consistently outperform single-modality approaches, *e.g.,* Miniotas *et al.* [33] and Zhang *et al.* [51].

EXPLORATORY STUDY

To understand the implications of using gaze+voice augmented assistants in everyday scenarios, we devised a Wizard-of-Oz experiment. This allowed us to gather user feedback on the use of WorldGaze against competitor techniques, without implementation limitations.

Setup

As an exemplary task, we asked participants to retrieve six pieces of information (*e.g.,* opening hours, ratings, phone numbers) about five restaurants within view on a busy commercial street. Participants completed this task three times, using one of three randomly ordered (Latin Square) METHODS: *Touch*, *Voice*, and *WorldGaze*. In the *Touch* condition, we asked participants to use Google Maps to query information (app already open on the smartphone). In the *Voice* condition, we used a Wizard-of-Oz voice assistant (triggered by "Hey Siri") that always returned the correct answer. Finally, in the *WorldGaze* condition, the voice assistant similarly returned the correct answer. Gaze was not tracked, but the experimenter asked participants to look at the restaurant in question while inquiring. For all methods, question order was randomized, with the added constraint that the same restaurant was never the target twice in a row.

Procedure

After welcoming participants, we explained the study, answered all open questions, and then asked them to give informed consent. We then went through the three conditions on the street in Latin Square order. After each condition, we asked participants to fill out a System Usability Scale (SUS) [11] (10-items on a 5-point Likert scale) and a raw NASA TLX questionnaire [22] (6-items on a 21-point Likert scale) and a single question on future use desirability (7-point Likert scale). Lastly, we conducted an exit interview capturing general feedback.

Participants

We recruited 12 participants (9 male and 3 female) from our institution with a mean age of 25.5 years (SD = 3.3). For this study, we only recruited participants with at least introductory coursework in HCI. The study took roughly 30 minutes, and participants were compensated $10 for their time.

Quantitative Feedback

After calculating the SUS score [11] between 0 and 100, we ran a Shapiro-Wilk normality test. As $p < .003$, we performed a Friedman test revealing a significant effect of METHOD on SUS rating ($\chi2(2) = 6.000$, $p = .0498$, $\eta^2 = .75$); see Figure 2 left. As post-hoc tests, we performed Wilcoxon rank sum test with Bonferroni correction. However, the post-hoc tests did not reveal any significant difference ($p > .05$).

After calculating the raw NASA TLX score [22], we ran a Shapiro-Wilk normality test. With $p < .002$, we performed an additional Friedman test for raw TLX revealing a significant effect of METHOD on raw TLX rating ($\chi2(2) = 7.787$, $p = .020$, $\eta^2 = .45$); see Figure 2 center. For post-hoc tests we performed Wilcoxon rank sum test with Bonferroni correction and found that Touch had a significantly higher task load than Voice ($p = .039$); all other combinations ($p > .05$). For Future Use we also ran a Shapiro-Wilk normality test ($p < .001$). Thus, we performed a third Friedman, which revealed no significant effect of METHOD on Future Use ($\chi2(2) = 1.929$, $p = .381$, $\eta^2 = .08$); see Figure 2 right.

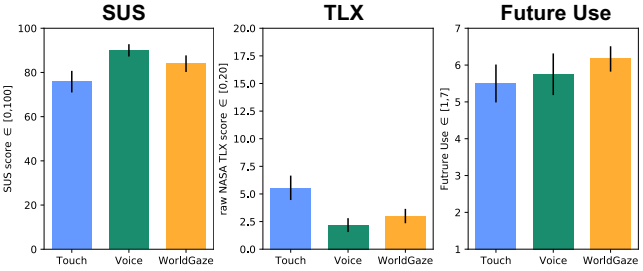

Figure 2. Left: System Usability Scale for our three conditions (lower is better). Center: raw NASA TLX rating (lower is better). Right: Rating of future use desirability (high is better).

As Shapiro-Wilk normality showed the normal distribution was not violated ($p > .052$), we performed a one-way ANOVA, which revealed that there was no significant effect of METHOD on Task Completion Time (TCT) ($F(2,22) = 0.013$, $p = .987$, $\eta^2 < .01$), with Touch ($M = 7.9$ sec, $SD = 4.1$), Voice ($M = 7.9$ sec, $SD = 4.3$), and WorldGaze ($M = 8.2$ sec, $SD = 4.9$). However, as WorldGaze requires less words to be articulated, utterance duration is shorter.

Qualitative Feedback

One researcher transcribed the interviews ($M = 15$min) [9], and then two researchers on the team affinity diagrammed [21] printed quotes to identify high-level themes, which we now summarize:

Easy and Natural: Most participants found WorldGaze to be a natural interaction (P1,2,4,6-9,11). Eleven participants (P1-7,8-12) articulated that WorldGaze is easy to use, "*implicit input with [WorldGaze] would be striking*" (P9) and "*a very discreet way to get information*" (P8). For example, P3 said WorldGaze is "*providing a more intuitive and real-time detailed inquiry*" and P4 stated that "*gaze is more socially acceptable*".

Useful and Fast: Five participants (P5,7,9,12) expressed that "*[WorldGaze] would be useful to have*" P5. Eight participants (P1-3,5-7,10,11) saw utility in the possibility of disambiguating between objects and places, for instance, P7 said "*I feel like inputting the gaze will help solve some of the accuracy problems that make voice assistants unreliable.*" Participants also identified that WorldGaze is useful for situations where the name of the object/place is unknown (P3,5,7,11). Additionally, six participants (P1-3,5,7,8) commented on the speed of WorldGaze, noting that touch felt slower: "*[WorldGaze] is faster - or it feels faster anyway - less frustrating*" (P2).

Novelty and Usability: As one would expect with a new input modality, several participants stated that they would have

to get used to WorldGaze before feeling comfortable (P2,3,8, 10,11). We also received feedback that WorldGaze required the user to hold the phone fairly high (P1,2,8-12). Five participants (P3,4,6,8,9) expressed concerns about accessibility ("*[people with] low vision*" P8) and social acceptance ("*people may think I'm recording them*" P4). We also received feedback on feasibility, with participants stating that WorldGaze may not work for places that are far away (P2-4,7,9,11), objects which are too close (P1,6,7), and that the latest generation of phones would be needed (P1,2,4).

Use Scenarios: Participants envisioned many uses for WorldGaze, including asking questions about products in stores or menu items in restaurants (P6,7,9,10,12). Interacting with smart home objects, such as controlling the TV or lighting, was mentioned by four participants (P5,6,7,10). Also mentioned were use cases in museums (P4,8), navigation support (P2,9), and for desktop computer interaction (P3,5), *e.g.,* MAGIC pointing [50].

Enhanced Feedback: Seven participants expressed a desire for better feedback in WorldGaze (P1,2,4,5,7,9,10), for example, an indicator that WorldGaze had selected the correct target (*e.g.* displaying a map or image of the restaurant). Six participants proposed improvements (P1,3,7,9,11), including an overlay on the camera view (*e.g.,* outline on the place of interest). In cases where the system selected the wrong location, participants proposed various resolution strategies, including giving multiple options based on the likelihood, using mid-air gestures, and a mode where the current gaze target was announced out loud. Finally, P8 mentioned a desire to use WorldGaze in concert with silent speech [46] and also conventional touch interaction.

New Interactions: Six participants (P1,4,6,7,11,12) suggested the system could be integrated into smart glasses ("*the most frictionless option*" P11) or added to camera-equipped smart devices (*e.g.,* Facebook Portal, Google Nest Hub). Another feature envisioned was to use WorldGaze to rapidly compare multiple objects or places (P2,5,9). Finally, participants suggested that WorldGaze could be a proactive system (P2,4,8), wherein a virtual assistant knows a user's focus and "*could make recommendations*" (P3) on the go.

IMPLEMENTATION

Our exploratory study gave us confidence that our technique would be quick, natural, and appreciated by users. The next challenge was figuring out how to create such an interaction technique without having to instrument the user or environment in any manner, and ideally, use only sensors already present in contemporary smartphones. We decided on a camera-only approach, taking advantage of recent trends in mobile hardware.

Platform Selection

At the time of writing, only iOS 13.0 and later permitted front and back cameras to be opened simultaneously, and it is for this reason that we selected iPhones as the platform for our proof-of-concept implementation. That said, this is not an innate hardware limitation; Android devices could have similar capabilities in the near future.

Device & Field of View

We used an iPhone XR for development and testing. This has a rear 12MP camera with a 67.3° field of view (FOV) and a 7MP front-facing camera with a 56.6° FOV. We note this FOV is considerably narrower than the most recent generation of flagship phones, including the Galaxy S10 series at 123°, iPhone 11 at 120°, Huawei P30 Pro at 120°, Asus ROG at 120°, and OnePlus 7 Pro at 117°. For front-facing (*i.e.,* "selfie") cameras, higher-end models often feature a FOV of around 90° (*e.g.,* Pixel 3 at 97°, and LG V40 at 90°), which we found to be more than sufficient to fully capture the head, even at closer ranges, such as when reading the screen. This increased FOV trend looks set to continue, and over time, one can expect these high-end camera features to trickle down to mid-tier phones, especially if there were additional driver applications such as WorldGaze.

We also note that with techniques such as visual odometry and SLAM [8] [17] – like that employed in Apple's ARKit – an object could still be addressed with gaze even if it is not currently seen in the rear camera view. Instead, the gaze vector could be projected into a 3D scene model stored in memory to much the same effect.

Head Gaze Ray Casting

Having selected iOS as our development platform, we could also leverage capabilities provided by the ARKit 3 SDK. This includes a robust face API offering six-degree-of-freedom tracking using the front-facing camera. We use the forward-facing head vector (GazeVector) to extend a ray out from the bridge of the nose, which we then project into the scene captured by the rear-facing camera. This vector is used in subsequent processes, such as performing hit testing with elements in the world (*e.g.,* restaurants or smart home devices). On an iPhone XR, this process runs at 30 FPS with ~50 ms of latency.

Object Recognition & Segmentation

A raycast into a scene denotes an area of interest, but it does not immediately provide a well-defined item of interest. Some items are large (*e.g.,* restaurant facade), while others are small (*e.g.,* bus stop sign). It may be that a user is looking at a menu on a restaurant window vs. the restaurant as a whole, also suggesting a hierarchy of foci. Thus, a parallel process is needed to resolve a user's true gaze intent, which then serves as an input parameter to *e.g.,* voice assistants.

Most straightforward is to use vision-based object recognition systems, such as Yolo [39], Google Vision [19], RetinaNet [29], and DenseCap [26], which provide bounding boxes. Even tighter semantic segmentation can be achieved with pipelines such as SegNet [5] and Mask R-CNN [23], which provide object contours. Although default models generally provide only generic class names (*e.g.,* "Car", but not "2019 Honda Civic"), they can also be trained to recognize specific object if given sufficient data. For example,

many mobile AR SDKs (*e.g.,* Vuforia [48]) allow developers to preregister specific objects and places for later recognition, and this is the approach we foresee in a commercial implementation of WorldGaze. There could also be a cloud-mediated library where *e.g.,* brick and mortar businesses and consumer goods manufacturers register their storefronts and wares.

As a proof of concept, we use Apple's Vision Framework [3] for object recognition and tracking. This API allows developers to register both 3D objects (*e.g.,* cars, appliances and furniture via the ARReferenceObject API), as well as planar images (*e.g.*, business logos and street signage via the ARReferenceImage API). We chose this over other similar packages chiefly for its excellent performance on the iPhone XR (hardware accelerated using Apple's A12 Bionic chip), allowing our whole stack to run at camera frame rate.

For each frame, we rank order all identified targets by confidence, using the minimum 3D distance of the gaze ray to the centroid of the object, weighted by the size of the object. The latter helps improve robustness in the case of nested objects. A fully probabilistic approach could also be powerful, leveraging frameworks that handle inputs with uncertainty [43].

Voice Assistant Integration

The final piece of our full stack implementation is integration with a voice agent. For this, we start by using the continuous listening feature on iOS combined with speech-to-text [4]. More specifically, we listen for "*Hey Siri*" as a keyword to start transcription of a voice query. We then search this text string for ambiguous nouns (*e.g.*, "*this*" and "*that place*"), replacing instances with the name of the object with the highest gaze probability (see previous section). We note that more advanced NLP methods could handle more complex phrasings, but our search and replace approach was sufficient as a prototype. In a commercia implementation, the updated phrase would be pushed back into the conventional voice assistant stack. However, to better control the user experience for testing and demonstration, we constrain the possible answers using a query-reply lookup table.

Comparative Approaches

Voice-only query approaches require users to be very explicit in defining objects or places of interest. At the time of writing, we found that even when standing directly in front of a Starbucks, asking Apple's Siri "when does Starbucks close?" required an additional voice step of confirming the Starbucks nearest to the user; see Video Figure. In general, geolocation technologies like GPS and WiFi positioning are too coarse for selecting individual storefronts, and of course, you often wish to inquire about something across the street or ahead of you. Indoors, you might wish to specify something as small as a thermostat in a dense scene of potential target objects. As before, voice is more useful for interacting with objects father away, not ones directly in front of you, where touch input might be more effective, and thus even centimeter indoor location in not a panacea for ambiguous voice queries.

Of course, WorldGaze is not the only option for specifying a distant, yet well-defined target without explicit speech. For example, instead of looking at a target, one could orient their phone towards it, which is how most mobile AR applications work today. While certainly more practical, it has the downside of having to "live through your phone" and makes rapid, ad hoc inquiries harder – one would have to launch the pass-through AR app to specify the target with any degree of accuracy. Another option is pointing with the hands [10], though this generally requires precise motion tracking [31] [32] and currently generation phones do not capture the hands unless they are fully extended outwards or held in front of the head, which is hardly natural.

Battery Life Implications

Although WorldGaze could be launched as a standalone application, we believe it is more likely for WorldGaze to be integrated as a background service that wakes upon a voice assistant trigger (*e.g.,* "Hey Siri"). Although opening both cameras and performing computer vision processing is energy consumptive, the duty cycle would be so low as to not significantly impact battery life of today's smartphones. It may even be that only a single frame is needed from both cameras, after which they can turn back off (WorldGaze startup time is 7 sec). Using bench equipment, we estimated power consumption at ~0.1 mWh per inquiry.

EVALUATION

We conducted a second study to evaluate the tracking and targeting performance of our proof-of-concept WorldGaze implementation.

Setup

In this study, participants were asked to stand in front of a wall at three different distances (DISTANCE: 1m, 2m, and 4m) while holding a phone and pointing with head gaze at 15 different targets (TARGET). The targets where arranged in a 5×3 grid with a center-to-center spacing of 80cm; Figure 4. Each target was registered as a separate object in the phone's WorldGaze database. The order of the three DISTANCE conditions was balanced using a Latin Square design, while the order of the targets was fully randomized (repeated three times each).

Procedure

After welcoming participants, we explained the study procedure and answered any questions. We then familiarized participants with the WorldGaze technique. Importantly, we gave no feedback (visual or otherwise) of the gaze ray to participants so as to not influence their targeting behavior. Gaze targets were announced one at time by the experimenter. Participants verbally announced (*e.g.,* "ok") when they were looking at the requested target, and the experimenter pressed a space bar on a laptop study interface, which recorded the selected object as reported by WorldGaze running on the phone, as well as accessory information for later analysis like the gaze vector. The next trial began automatically.

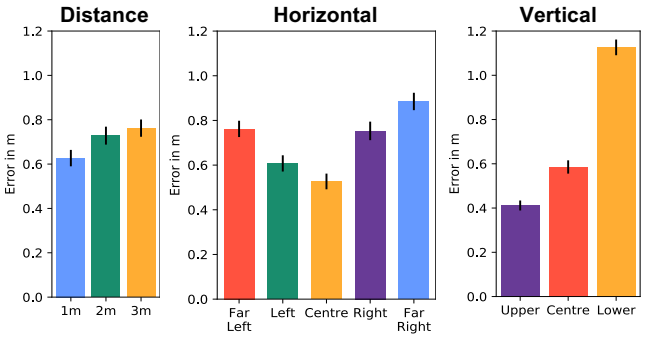

Figure 3. Left: error vs. user distance from wall. Center: error with respect to target horizontal placement. Right: error based on target vertical placement.

Participants

We recruited 12 participants (9 male, and 3 female) from our institution with a mean age of 28.9 years (SD = 5.8). The only requirement was that they had no locomotor impairment. The study took approximately 20 minutes and participants were compensated $10 for their time.

Results

In total, participants gaze-selected 1620 targets (12 participants, 5×3 grid of targets, 3 repeats per target, and 3 distance conditions). Overall, across all conditions and participants, we found a mean tracking error in real-world coordinates of 0.71m (SD = 0.47). Note that this result is cross-user (*i.e.,* "out of the box" accuracy), with no per-user or post hoc global corrections. We first processed the tracking data so that the grid aligned from all sessions. In line with prior related work, we filtered outlier trials with error exceeding mean+3SD [31] [32], which removed 11 targets. The mean error was lower when standing close to the wall, and highest when farther away.

A Shapiro-Wilk normality test showed that the Error is not normally distributed ($p < .038$), thus, we performed a three-way ART RM-ANOVA [49]. The analyses revealed a statistically significant influence of DISTANCE on Error ($F(2,483.0) = 19.6$, $p < .001$); see Figure 3 left. When

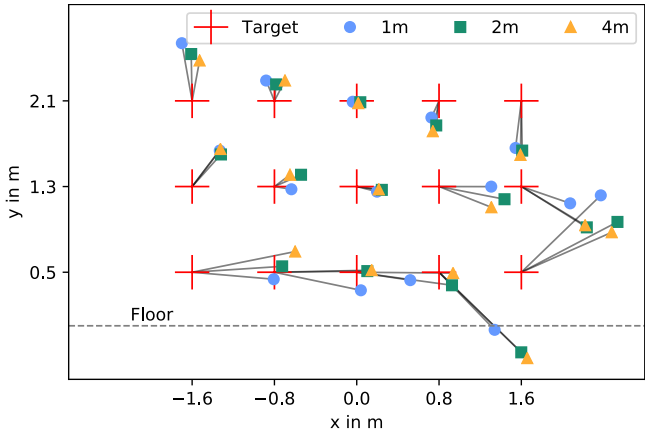

Figure 4. Mean gaze error for the 15 targets at three distances. Error is highest in the lower right corner, where participants had to look over their right arm to see the target.

breaking error out by HORIZONTAL and VERTICAL accuracy, we find statistically significant impact on Error ($F(4,483.0) = 50.8$, $p<.001$; $F(2,483.0) = 521.5$, $p < .001$; respectively); see Figure 3 center and right. Further we found that all two-way interaction effects are significant ($p < .002$), but none of the three-way interactions ($p = .755$).

The center column was the least error-prone (Figure 3 center and right), perhaps because it was always closest and it is easier to look straight ahead. Our results also show that targets situated higher are more precise, which is advantageous since most foci of interest in outdoor scenes are located at eye height or above (*e.g.,* signage). Finally, Figure 4 shows an overall correlation between target placement and error, which is in line with errors shown for traditional mid-air pointing using head-finger raycasting (*c.f.,* Mayer *et al.* [31][32]). Overall, we are confident that our current implementation could be used to select small but sparse objects, such as a lamp on a table, and is certainly accurate for most outdoor uses.

EXAMPLE USES

We now briefly describe example interactions in three use domains where we believe WorldGaze could be particularly useful: streetscapes, smart homes/offices, and retail. Please also see our Video Figure for a real-time demonstration of our implementation.

Streetscapes

It is not uncommon to see people walking down the street looking at their smartphones; see Figure 1, left. With sufficiently wide-angle lenses, WorldKit could allow for natural, rapid, and targeted voice inquiries. For example, a user could look at a store front and ask, "when does *this* open?" WorldGaze fills in the ambiguous "*this*" with the target business, allowing the voice agent to reply intelligently. Similarly, a user could ask "what is the rating for *this* place" or even "make me a reservation for 2 at 7pm"; see Figure 5.

Retail

Retail settings are also ripe for augmentation, as they are full of a great variety of objects that customers might wish to know more information about; see Figure 6. For example, a customer could ask, "does *this* come in any other colors?" in regard to a sofa they are evaluating. Likewise, they could also say "add *this* to my wishlist". It would also be trivial to extend WorldGaze to handle multiple sequential targets, allowing for comments such as "what is the price difference between *this*… and *this*."

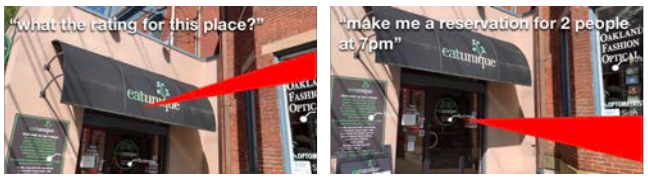

Figure 5. WorldGaze, in concert with a voice agent, could enable much more natural and rapid retrieval of information about e.g., businesses while walking down a street.

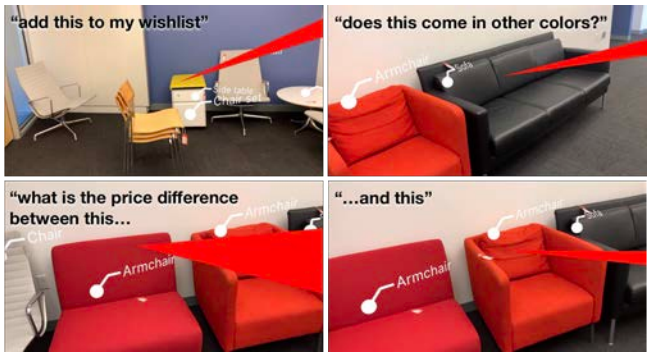

Figure 6. In retail settings, WorldGaze-augmented shopping apps could allow users to rapidly retrieve item information. We also implemented an example interaction of a user specifying two targets in one voice inquiry.

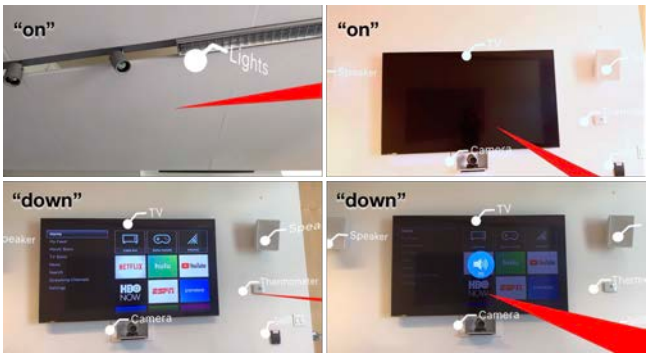

Figure 7. WorldGaze could be especially useful in settings with many IoT appliances, where extra context could be used to resolve otherwise ambiguous verbs, like go, play or start.

Smart Homes and Offices

Finally, WorldGaze could also facilitate rapid interactions in density instrumented smart environments, automatically resolving otherwise ambiguous verbs, such as play, go, and stop; see Figure 7. For example, a user could say "on" to lights or a TV, or "down" to a TV or thermostat. WorldGaze offers the necessary context to resolve these ambiguities and trigger the right command; see Video Figure.

LIMITATIONS & FUTURE WORK

As noted previously, our current WorldGaze implementation is constrained by the rear camera's field of view – wider-angle lenses mean more of the world is gaze addressable. Fortunately, the current trend in smartphones is to include wide angle lenses, with some models exceeding 120°. While this falls short of human vision, with roughly a 135° horizontal FOV per eye [18], it is sufficient to capture the majority of a scene in front of a user. Overall, we foresee this FOV gap diminishing overtime, especially if capabilities such as WorldGaze are an additional driving factor.

That said, we note that a limited FOV might be partially overcome through future integration of techniques like visual odometry and SLAM [8] [17], which can iteratively build a 3D world scene in memory. As the smartphone's 3D position in the scene is known, along with the live head vector, user could gaze at previously captured objects with no difference in the interaction.

We also note that we started our implementation efforts utilizing both eye gaze and head orientation, which would provide a fine-grained gaze vector perfect for WorldGaze. We tested numerous state-of-the-art algorithms [6], [28], [35][52], but found accuracy to be severely lacking for our particular use case. WorldGaze operates at longer ranges than most screen-based gaze interactions, which exacerbates error; *e.g.,* ±15° angular error equates to meter-scale inaccuracies when looking at objects four meters away. Instead, we decided to build our proof-of-concept implementation on head gaze alone, which is more stable and accurate (chiefly because there are plenty of facial landmarks onto which to fit a 3D head model). Of course, aiming with one's head is less natural than gazing with the eyes, and so we are hopeful that eye tracking sensors and algorithms capable of running on mobile devices will continue to improve.

CONCLUSION

We have presented our work on WorldGaze, an interaction technique leveraging front and rear smartphone cameras that allows users to denote an object or region of interest with their head direction. With computer-vision-based object recognition, we can identify what *e.g.,* business or IoT device a user is looking at, which we can pass as extra physical context to voice agents like Siri and Alexa, making them considerably more natural and contextually aware. We show through qualitative and quantitative studies that such a feature would be welcomed by users and is accurate to around one meter in the world. Finally, as remarked by our participants, WorldGaze could prove valuable in form factors beyond smartphones, such as smart glasses, which we hope to explore in the future.

REFERENCES

[1] Cengiz Acartürk, João Freitas, Mehmetcal Fal, and Miguel Sales Dias. 2015. Elderly Speech-Gaze Interaction. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, Cham, 3-12. DOI: https://doi.org/10.1007/978-3-319-20678-3_1

[2] Lisa Anthony, Jie Yang, and Kenneth R. Koedinger. 2005. Evaluation of multimodal input for entering mathematical equations on the computer. In CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05). ACM, New York, NY, USA, 1184-1187. DOI: http://dx.doi.org/10.1145/1056808.1056872

[3] Apple Vision Framework. 2019. URL: https://developer.apple.com/documentation/vision

[4] Apple Speech Framework. 2019. URL: https://developer.apple.com/documentation/speech

[5] Vijay Badrinarayanan, Kendall Alex, and Cipolla Roberto. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE*

transactions on pattern analysis and machine intelligence 39.12: 2481-2495. DOI: http://dx.doi.org/10.1109/TPAMI.2016.2644615

[6] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision* (WACV '16). IEEE, 1-10. DOI: http://dx.doi.org/10.1109/WACV.2016.7477553

[7] Tanya R. Beelders, and Pieter J. Blignaut. 2011. The Usability of Speech and Eye Gaze as a Multimodal Interface for a Word Processor. *Speech Technologies,* 386-404. DOI: http://dx.doi.org/10.5772/16604

[8] Tim Bailey, and Hugh Durrant-Whyte. 2006. Simultaneous localization and mapping (SLAM): Part II. *IEEE robotics & automation magazine* 13, no. 3, 108-117. IEEE. DOI: http://dx.doi.org/10.1109/MRA.2006.1678144

[9] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI research: Going behind the scenes. Synthesis lectures on human-centered informatics, 9(1), 1-115. DOI: https://doi.org/10.2200/S00706ED1V01Y201602HCI034

[10] Richard A. Bolt. 1980. Put-that-there: Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques* (SIGGRAPH '80). ACM, New York, NY, USA, 262-270. DOI: http://dx.doi.org/10.1145/800250.807503

[11] John Brooke. 1996. SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7.

[12] Drini Cami, Fabrice Matulic, Richard G. Calland, Brian Vogel, and Daniel Vogel. 2018. Unimanual Pen+Touch Input Using Variations of Precision Grip Postures. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18). ACM, New York, NY, USA, 825-837. DOI: https://doi.org/10.1145/3242587.3242652

[13] Ishan Chatterjee, Robert Xiao, and Chris Harrison. 2015. Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15). ACM, New York, NY, USA. DOI: https://doi.org/10.1145/2818346.2820752

[14] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, and Benjamin Cowan. 2018. The State of Speech in HCI: Trends, Themes and Challenges. In *Proceedings of the Interacting with Computers*. DOI: https://doi.org/10.1093/iwc/iwz016

[15] Heiko Drewes, Alexander De Luca, and Albrecht Schmidt. 2007. Eye-gaze interaction for mobile phones. In *Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology* (Mobility '07). ACM, New York, NY, USA, 364-371. DOI: http://dx.doi.org/10.1145/1378063.1378122

[16] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2015. Orbits: Gaze Interaction for Smart Watches using Smooth Pursuit Eye Movements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (UIST '15). ACM, New York, NY, USA, 457-466. DOI: https://doi.org/10.1145/2807442.2807499

[17] Jorge Fuentes-Pacheco, José Ruiz-Ascencio, and Juan Manuel Rendón-Mancha. 2015. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review* 43, no. 1, 55-81. DOI: https://doi.org/10.1007/s10462-012-9365-8

[18] Alastair G. Gale. 1997. Human response to visual stimuli. In *The perception of visual information*. Springer, New York, NY, 127-147. DOI: https://doi.org/10.1007/978-1-4612-1836-4_5

[19] Google Cloud Vision AI. 2019. https://cloud.google.com/vision/automl/object-detection/docs/

[20] Floyd A. Glenn III, Helene P. Iavecchia, Lorna V. Ross, James M. Stokes, William J. Weiland, Daniel Weiss, and Allen L. Zaklad. 1986. Eyevoice-controlled interface. In *Proceedings of the Human Factors Society,* 322-326. DOI: https://doi.org/10.1177/154193128603000402

[21] Gunnar Harboe, and Elaine M. Huang. 2015. Real-World Affinity Diagramming Practices: Bridging the Paper-Digital Gap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, New York, NY, USA, 95–104. DOI: http://dx.doi.org/10.1145/2702123.2702561

[22] Sandra G. Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50, No. 9, 904-908, Los Angeles, CA, Sage publications. DOI: https://doi.org/10.1037/e577632012-009

[23] Kaiming He, Gkioxari Georgia, Dollár Piotr, and Girshick Ross. 2017. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*. IEEE, 2961-2969. DOI: http://dx.doi.org/10.1109/TPAMI.2018.2844175

[24] Ken Hinckley, Koji Yatani, Michel Pahud, Nicole Coddington, Jenny Rodenhouse, Andy Wilson, Hrvoje Benko, and Bill Buxton. 2010. Pen + touch = new tools. In Proceedings of the 23nd annual ACM symposium on User interface software and technology (UIST '10). ACM, New York, NY, USA, 27-36. DOI: https://doi.org/10.1145/1866029.1866036

[25] Ron Jacob. 1995. Eye tracking in advanced interface design. In *Virtual Environments and Advanced Interface Design.* New York: Oxford University Press, 258-288.

[26] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR '16). IEEE 4565-4574. DOI: https://doi.org/10.1109/CVPR.2016.494

[27] David B. Koons, Carlton J. Sparrell, and Kristinn R. Thorisson. 1993. Integrating simultaneous input from speech, gaze, and hand gestures. MIT Press: Menlo Park, CA, 257-276.

[28] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2016* (CVPR '16). IEEE, 2176-2184. DOI: https://doi.org/10.1109/CVPR.2016.239

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (CVPR '17). IEEE, 2117-2125. DOI: https://doi.org/10.1109/CVPR.2017.106

[30] Diako Mardanbegi, and Dan Witzner Hansen. 2011. Mobile gaze-based screen interaction in 3D environments. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications* (NGCA '11). ACM, New York, NY, USA, Article 2, 4 pages. DOI: http://dx.doi.org/10.1145/1983302.1983304

[31] Sven Mayer, Katrin Wolf, Stefan Schneegass, and Niels Henze. 2015. Modeling Distant Pointing for Compensating Systematic Displacements. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, New York, NY, USA, 4165-4168. DOI: https://doi.org/10.1145/2702123.2702332

[32] Sven Mayer, Valentin Schwind, Robin Schweigert, and Niels Henze. 2018. The Effect of Offset Correction and Cursor on Mid-Air Pointing in Real and Virtual Environments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18). ACM, New York, NY, USA, Paper 653, 13 pages. DOI: https://doi.org/10.1145/3173574.3174227

[33] Darius Miniotas, Oleg Špakov, Ivan Tugoy, and I. Scott MacKenzie. 2006. Speech-augmented eye gaze interaction with small closely spaced targets. In *Proceedings of the 2006 symposium on Eye tracking research & applications* (ETRA '06). ACM, New York, NY, USA, 67-72. DOI: http://dx.doi.org/10.1145/1117309.1117345

[34] Robert Neßelrath, Mohammad Mehdi Moniri, and Michael Feld. 2016. Combining speech, gaze, and micro-gestures for the multimodal control of in-car functions. In *Proceedings of the 12th International Conference on Intelligent Environments* (IE '16). IEEE. DOI: http://dx.doi.org/10.1109/IE.2016.42

[35] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. Webgazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016.*

[36] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, and Hans Gellersen. 2014. Gaze-touch: combining gaze with multi-touch for interaction on the same surface. In Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST '14). ACM, New York, NY, USA, 509-518. DOI: https://doi.org/10.1145/2642918.2647397

[37] Bastian Pfleging, Stefan Schneegass, and Albrecht Schmidt. 2012. Multimodal interaction in the car: combining speech and gestures on the steering wheel. In Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '12). ACM, New York, NY, USA, 155-162. DOI: http://dx.doi.org/10.1145/2390256.2390282

[38] Katrin Plaumann, Matthias Weing, Christian Winkler, Michael Müller, and Enrico Rukzio. 2018. Towards accurate cursorless pointing: the effects of ocular dominance and handedness. Personal Ubiquitous Comput. 22, 4 (August 2018), 633-646. DOI: https://doi.org/10.1007/s00779-017-1100-7

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (CVPR '16). IEEE, 779-788. DOI: https://doi.org/10.1109/CVPR.2016.91

[40] Florian Roider, Lars Reisig, and Tom Gross. 2018. Just Look: The Benefits of Gaze-Activated Voice Input in the Car. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (AutomotiveUI '18). ACM, New York, NY, USA, 210-214. DOI: https://doi.org/10.1145/3239092.3265968

[41] David Rozado, Alexander McNeill, and Daniel Mazur. 2016. Voxvisio – Combining Gaze And Speech For Accessible Hci. In *Proceedings of RESNA/NCART 2016.*

[42] Albrecht Schmidt, Michael Beigl, and Hans Gellersen. 1999. There is more to context than location. Computers & Graphics 23.6, 893-901. DOI: https://doi.org/10.1016/S0097-8493(99)00120-X

[43] Julia Schwarz, Scott Hudson, Jennifer Mankoff, and Andrew D. Wilson. 2010. A framework for robust and flexible handling of inputs with uncertainty. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (UIST '10). ACM, New York, NY, USA, 47-56. DOI: https://doi.org/10.1145/1866029.1866039

[44] Robin Schweigert, Valentin Schwind, and Sven Mayer. 2019. EyePointing: A Gaze-Based Selection Technique. In *Proceedings of Mensch und Computer 2019* (MuC '19). ACM, New York, NY, USA, 719-723. DOI: https://doi.org/10.1145/3340764.3344897

[45] Valentin Schwind, Sven Mayer, Alexandre Comeau-Vermeersch, Robin Schweigert, and Niels Henze. 2018. Up to the Finger Tip: The Effect of Avatars on Mid-Air Pointing Accuracy in Virtual Reality. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '18). ACM, New York, NY, USA, 477-488. DOI: https://doi.org/10.1145/3242671.3242675

[46] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (UIST '18). ACM, New York, NY, USA, 581-593. DOI: https://doi.org/10.1145/3242587.3242599

[47] Daniel Vogel, and Ravin Balakrishnan. 2005. Distant freehand pointing and clicking on very large, high resolution displays. In *Proceedings of the 18th annual ACM symposium on User interface software and technology* (UIST '05). ACM, New York, NY, USA, 33-42. DOI: http://dx.doi.org/10.1145/1095034.1095041

[48] Vuforia. URL: https://developer.vuforia.com

[49] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11). ACM, New York, NY, USA, 143-146. DOI: https://doi.org/10.1145/1978942.1978963

[50] Shumin Zhai, Carlos Morimoto, and Steven Ihde. 1999. Manual and gaze input cascaded (MAGIC) pointing. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '99). ACM, 246-253. DOI: http://dx.doi.org/10.1145/302979.303053

[51] Qiaohui Zhang, Atsumi Imamiya, Kentaro Go, and Xiaoyang Mao. 2004. Resolving ambiguities of a gaze and speech interface. In *Proceedings of the 2004 symposium on Eye tracking research & applications* (ETRA '04). ACM, New York, NY, USA, 85-92. DOI: https://doi.org/10.1145/968363.968383

[52] Xucong Zhang, Yusuke Sugano, M. Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (CVPR '15). IEEE, 4511-4520. DOI: https://doi.org/10.1109/CVPR.2015.7299081