Anhong Guo — Research Statement

As a human-computer interaction (HCI) researcher, I create hybrid human- and AI-powered intelligent interactive systems to provide access to visual information in the real world. By combining the advantages of humans and AI, these systems can be nearly as robust and flexible as humans, and nearly as quick and low-cost as automated AI, enabling us to solve problems that are currently impossible with either alone.

In my dissertation work, I developed and deployed human-AI systems for two application domains: accessibility and environmental sensing. To make physical interfaces accessible for blind people, I developed systems to interpret static and dynamic interfaces, enabling blind people to independently access them through audio feedback or tactile overlays [1,2,3]. For environmental sensing, I developed and deployed a camera sensing system that collects human labels to bootstrap automatic processes to answer real-world visual questions, allowing end users to actionalize AI in their everyday lives [4]. AI systems often require huge amount of up front training data to get started, but targeted human intelligence can bootstrap the systems with relatively little data. Although humans may be slower initially, quickly bootstrapping to automated approaches provides a good balance, enabling human-AI systems to be scalable and rapidly deployable.

The goal of my research is to create intelligent interactive systems that solve AI-hard real-world problems. These systems collect data for users' immediate needs, in order to build a model to work in the moment. To the end users, these systems are always intelligent and smart. But under the hood, large-scale data can be collected, and automation can be achieved over time to support these user needs.

Human-AI Systems to Make Physical Interfaces Accessible

The world is full of physical interfaces that are inaccessible to blind people, from microwaves and information kiosks to thermostats and checkout terminals. The VizWiz dataset [7] I helped release showed that many blind people sought assistance using such interfaces. Blind people cannot access these interface because the buttons are tactually indistinguishable, and the screens contain visual information that they cannot read. Creating new devices that are accessible could work, but is unlikely to make it into all devices produced due to cost, let alone the substantial legacy of inaccessible devices already in the world.

To make physical interfaces accessible, I built *VizLens* [1], a robust and interactive screen reader for real-world static interfaces (Figure 1). To work robustly, VizLens combines on-demand crowdsourcing and real-time computer vision. When a blind person encounters an inaccessible interface for the first time, they use a smartphone camera to capture a picture of the device and then send it to the crowd. This picture then becomes a reference image. Within a few minutes, crowd workers mark the layout of the interface, annotate its elements (e.g., buttons or other controls), and describes each element. Later, when the person wants to use the interface, they open the VizLens application, point it towards the interface, and hover a finger over it. VizLens uses SURF-based object matching techniques to match the crowd-labeled reference image to the image captured in real-time, and track the user's finger to retrieve and provide audio feedback and guidance. Deep CNNs may increase the robustness, but the beauty of our approach is that even simple computer vision techniques work. With such instantaneous feedback, VizLens allows blind users to interactively explore and use inaccessible interfaces. VizLens trades off the advantages of humans and computer vision to be nearly as robust as a person in interpreting the interface and nearly as quick and low-cost as a computer vision system to re-identify the interface and provide real-time feedback. I further explored cursor-based interactions [5] at Google, integrating VizLens' real-world screen reader interaction as a type of finger cursor.

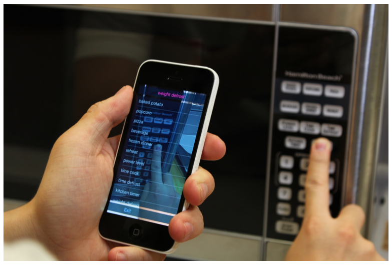
Figure 1: VizLens is a screen reader to help blind people access static physical interfaces.

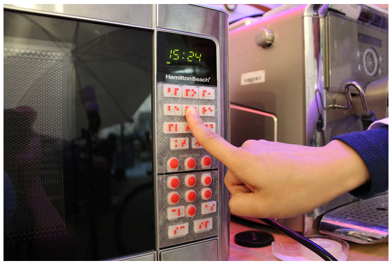

Figure 2: Facade enables blind users to independently create 3D-printed tactile overlays for appliances.

Blind people often label home appliances with Braille stickers, but doing so generally requires sighted assistance. I developed *Facade* [2], a crowdsourced fabrication pipeline that enables blind people to independently create 3D-printed tactile overlays for inaccessible appliances (Figure 2). Blind users capture a photo of an inaccessible interface with a readily available fiducial marker for recovering size information using perspective transformation. This image is then labeled by crowd workers. Facade then generates a 3D model for a layer of tactile and pressable buttons that fits over the original controls, which the blind users can customize using the iOS app. Finally, a home 3D printer or commercial service can be used to fabricate the layer. We went through several design iterations to determine the most effective overlay design, material configuration, and printer settings to make the 3D-printed overlays easy to attach, read, and press. Facade makes end-user fabrication accessible to blind people, by shifting the sighted assistance to a virtual crowd working with computer vision. Facade combines a human-AI interpretation pipeline with an accessible 3D printing application.

VizLens and Facade enable blind users to access many *static* interfaces. To make *dynamic* touchscreens such as public kiosks and payment terminals accessible, I then developed *StateLens*, a three-part reverse engineering solution [3]. First, using a hybrid crowd-computer vision pipeline (Figure 3a), StateLens reverse engineers the underlying state diagrams of existing interfaces using point-of-view videos found online or taken by users. Second, using the state diagrams, StateLens automatically generates conversational agents to guide blind users through specifying the tasks that the interface can perform, allowing the StateLens iOS application to provide interactive guidance and feedback so that blind users can access the interface. Finally, to address the "Midas touch problem" of accidental triggers during exploration, we designed a set of 3D-printed accessories (Figure 3b: finger cap and stylus) that allow users to explore without touching the screen, and perform a gesture to activate touch at a desired position. These accessories bring "risk-free exploration" to public capacitive touchscreens without modifying the underlying hardware or software, which is core to accessible touchscreen interaction. Our technical evaluation with 12 touchscreen devices and over 70K video frames showed that StateLens can accurately reconstruct interfaces from stationary, hand-held, and web videos; and through a user study with 14 blind participants, we showed that the complete system enables blind users to access otherwise inaccessible dynamic touchscreens.

StateLens addresses the very hard case in which blind users encounter a touchscreen in the real world that is inaccessible, which they cannot modify the hardware or software, and whose screen updates dynamically to show new information and interface components. Furthermore, StateLens takes advantage of different kinds of human intelligence: humans who provide access and collect videos at the interface to build up the training data, and on-

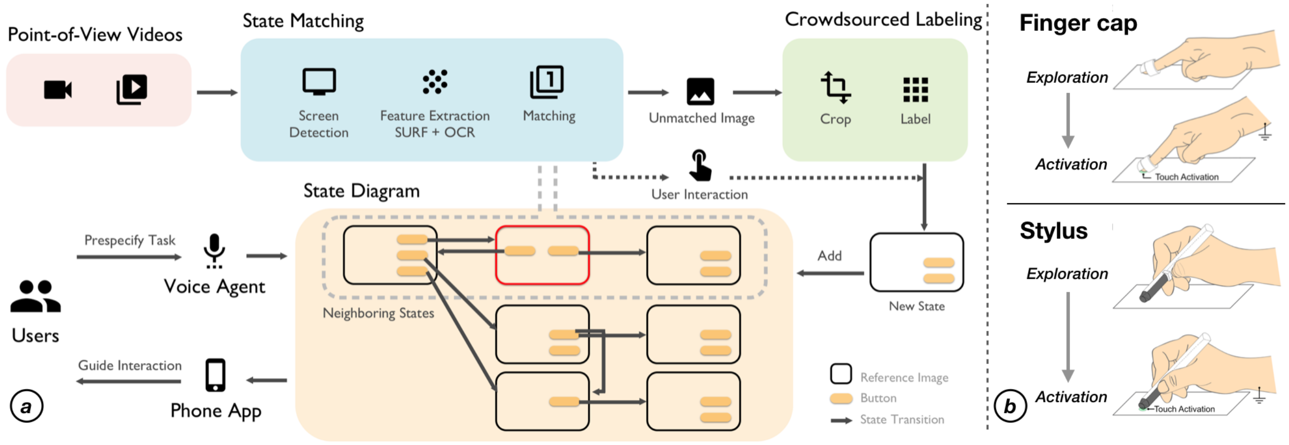

Figure 3: StateLens uses a hybrid crowd-computer vision pipeline to dynamically generate state diagrams about interface structures from point-of-view usage videos, and using the diagrams to provide interactive guidance and feedback to help blind users access the interfaces (a). 3D-printed accessories enable "risk-free exploration" (b).

line crowds who provide necessary labels to bootstrap automation. This line of work has been included as reading materials in several graduate human-computer interaction, crowdsourcing and accessibility classes, including at University of Washington, University of Maryland at College Park, and KAIST. Currently, I am working on deploying VizLens, in order to benefit blind people in the wild, and to collect a dataset for interfaces and interactions.

Human-AI Systems for Environmental Sensing

Beyond enabling access to physical interfaces for blind people, I also explored environmental sensing platforms for understanding the visual world. I developed and deployed *Zensors*++ [4], a human-AI camera sensing system to answer natural language user questions based on camera streams. To create a sensor, users select a camera, drag a bounding box to select a region of interest, and ask a natural language question. At first, crowd workers provide near-instant answers for users' questions. Over time, Zensors++ relies on the crowd less, as answers can be automated through perceptual image hashing and continuously-evolving machine learning. We deployed Zensors++ in the wild, with real users, over many months and environments, generating 1.6 million answers for nearly 200 questions created by our participants, costing roughly 6/10ths of a cent per answer delivered. We demonstrated that crowd workers were able to provide labels quickly (~6s) and at scale, and that the system could hand-off to image hashing and machine learning classifiers for ~75% of all questions.

Participants created a wide range of sensors for their use cases (Figure 4). For example, for "Are the tables set up in rows?", the instructor used it to decide whether he needed to go to the classroom early to arrange the room before lecture. For "Is someone sitting on this furniture?", the program director was using Zensors++ to conduct physical A/B testing on different furniture arrangements. For "Is the trash can full?", the building manager was able to get email notifications when the trashcan is full, so he could better allocate resources to clean them up, rather than doing periodic checking manually. For "How many people are in line at the cash register?", a restaurant manager was interested in using the longitudinal data to identify consumption patterns to better plan and prepare food, while students and faculties were more interested in knowing how long the line is. Overall, our deployments demonstrated that human-AI, camera-based sensing can work at scale. Zensors++ relies on end users to define questions of interest and specify image region, as well as online crowd workers to provide labels when necessary. Then it relies on machine intelligence to automate over time to reduce the cost and latency. This project has resulted in the startup Zensors Inc. (zensors.com), and is predicting wait times for Pittsburgh International Airport (news), tracking parking usage for Pittsburgh Parking Authority, and monitoring resource utilization of co-working spaces. I have also been recognized with several entrepreneurial and invention awards, including the CMU Swartz Innovation Fellowship, and the McGinnis Venture Capital Award.

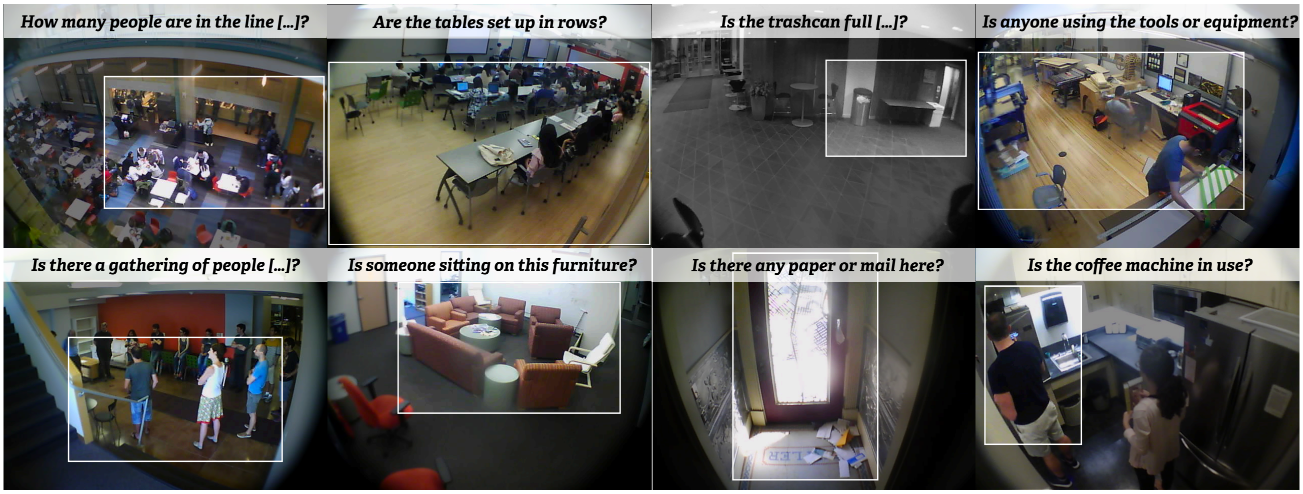
How many people are in the line [...]?
Are the tables set up in rows?
Is the trashcan full [...]?
Is anyone using the tools or equipment?
Is there a gathering of people [...]?
Is someone sitting on this furniture?
Is there any paper or mail here?
Is the coffee machine in use?

Figure 4: Eight example sensors created by our participants using Zensors++, with regions of interest highlighted on the full camera image. Many sensors directly complemented and augmented people's existing work practices.

Research Agenda

VizLens illustrates the approach I take in my research: I start by identifying a real problem (physical interfaces are not accessible), next understand where humans and machines work best for solving this problem (human for interpreting arbitrary interface, machine for remembering patterns and re-identifying later), then design human-AI systems as technology enablers, and finally deploy them in the wild to collect big data for understanding their limits and contributing back to the AI community for approaching automation via datasets (e.g., [7,8]). Here I outline opportunities that I am excited to pursue:

Accessibility as a Driving Force of AI. Accessibility is a unique problem domain because of its challenging constraints, but also the high potential value of technology enablers for end users. In my research, I have pushed boundaries in assistive technologies to enable visual access in both the real world [1,2,3,5] and the digital world [6]. Moving forward, I am excited to continue working with people with disabilities to solve real challenging problems people face, and in turn study their use to inform the directions of building better and beneficial AI.

AI Datasets and Fairness. In addition to developing system, pushing the boundaries of AI also requires better datasets rooted in human problems. I have collaborated with AI researchers in developing datasets for visual question answering [7] and privacy [8], and I plan to continue this direction, such as to collect a dataset of interfaces and interface interactions from the VizLens deployment. Relatedly, huge challenges exist in ensuring that the systems we are developing are fair for everyone, regardless of their gender, race, and disabilities. I have started to explore this direction by 1) proposing a roadmap for addressing fairness issues of AI systems for people with disabilities [9], 2) developing benchmarking datasets for revealing and auditing bias, and 3) investigating techniques for harvesting collective intelligence to uncover "unknown unknowns."

Designing Social and AR Systems. As I have shown in my research, the humans in human-AI systems are key to the success. The humans involved are also diverse: the end user, other users, and crowd workers. As I move forward, I plan to explore broader types of social systems integrating the advantages of humans and AI, and tightly weaving the physical and digital worlds. Along this direction, I have designed tools and systems allowing people to create together in augmented reality (AR) [10], write together on smartwatches [11], and reuse 3D models across users [12]. In the future, how can we create a virtual layer of hybrid intelligence, embedded in the physical world, and contributed by both humans and machines?

Enabling Access in More Contexts. The systems that I create could find broader applications in other domains. For example, StateLens could augment how people generally interact with touchscreen interfaces in the real world as cognitive assistance. When configuring complicated medical devices, or when interacting with machines in different languages, StateLens could provide guidance through visual overlays in AR. Furthermore, Zensors++ could assist blind users in sensing visual changes in their environments. My work on enabling no-touch, wrist-only interactions on smartwatches [13] has broader impact for not only people with situational impairments, but also for people with limb differences. Techniques for identifying user handprints on capacitive touchscreens [14] and presenting picking orders on head-up displays [15] could inform how assistive technologies be developed with limited hardware capabilities and with users' limited attention. In my future research, I will continue thriving to develop solutions that are generalizable across domains and contexts.

Collaboration. In graduate school, I have been fortunate to work with over 80 collaborators from over 20 institutions (including 5 companies). My research has benefited greatly from these collaborations, and I look forward to continuing this tradition as I move my research forward.

In summary, I am a technical HCI researcher who looks to identify real-world problems that people face in their everyday lives, then create intelligent systems that combine human and machine intelligence in new ways to solve the problems, and finally deploy them to collect data for understanding their usage in-the-wild. My broad background in HCI and computer science allows me to easily collaborate with people across many areas.

REFERENCES

[1] **Anhong Guo**, Xiang 'Anthony' Chen, Haoran Qi, Samuel White, Suman Ghosh, Chieko Asakawa, Jeffrey P. Bigham. VizLens: A Robust and Interactive Screen Reader for Interfaces in the Real World. In *Proceedings of the 29th Annual ACM Symposium on User Interface Software & Technology (UIST 2016)*. Tokyo, Japan. 2016.

[2] **Anhong Guo**, Jeeeun Kim, Xiang 'Anthony' Chen, Tom Yeh, Scott Hudson, Jennifer Mankoff, Jeffrey P. Bigham. Facade: Auto-generating Tactile Interfaces to Appliances. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2017)*. Denver, CO. 2017.

[3] **Anhong Guo**, Junhan Kong, Michael Rivera, Frank F. Xu, Jeffrey P. Bigham. StateLens: A Reverse Engineering Solution for Making Existing Dynamic Touchscreens Accessible. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software & Technology (UIST 2019)*. New Orleans, LA. 2019.

[4] **Anhong Guo**, Anuraag Jain, Shomiron Ghose, Gierad Laput, Chris Harrison, Jeffrey P. Bigham. Crowd-AI Camera Sensing in the Real World. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (Ubicomp 2018)* 2.3: 111. Singapore. 2018.

[5] **Anhong Guo**, Saige McVea, Xu Wang, Patrick Clary, Ken Goldman, Yang Li, Yu Zhong, Jeffrey Bigham. Investigating Cursor-based Interactions to Support Non-Visual Exploration in the Real World. In *Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2018)*. Galway, Ireland. 2018.

[6] Sujeath Pareddy, **Anhong Guo**, Jeffrey P. Bigham. X-Ray: Screenshot Accessibility via Embedded Metadata. In *Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2019)*. Pittsburgh, PA. 2019.
Best Artifact Award.

[7] Danna Gurari, Qing Li, Abigale Stangl, **Anhong Guo**, Chi Lin, Kristen Grauman, Jiebo Luo, Jeffrey P. Bigham. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*. Salt Lake City, Utah. 2018.
Spotlight Presentation.

[8] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, **Anhong Guo**, Abigale Stangl, Jeffrey P. Bigham. VizWiz-Priv: A Dataset for Recognizing the Presence and Purpose of Private Visual Information in Images Taken by Blind People. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*. Long Beach, CA. 2019.

[9] **Anhong Guo**, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, Meredith Ringel Morris. Toward Fairness in AI for People with Disabilities: A Research Roadmap. In *ACM ASSETS 2019 Workshop on AI Fairness for People with Disabilities (ASSETS 2019 AI Fairness Workshop)*. Pittsburgh, PA. 2019.

[10] **Anhong Guo**, Ilter Canberk, Hannah Murphy, Andrés Monroy-Hernández, Rajan Vaish. Blocks: Collaborative and Persistent Augmented Reality Experiences. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (Ubicomp 2019)* 3.3: 83. London, United Kingdom. 2019.

[11] Michael Nebeling, Alexandra To, **Anhong Guo**, Adrian de Freitas, Jaime Teevan, Steven Dow, Jeffrey P. Bigham. WearWrite: Crowd-Assisted Writing from Smartwatches. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI 2016)*. San Jose, CA. 2016.

[12] Jeeeun Kim, **Anhong Guo**, Tom Yeh, Scott Hudson, Jennifer Mankoff. Understanding Uncertainty in Measurement and Accommodating its Impact in 3D Modeling and Printing. In *Proceedings of the 2017 ACM Conference on Designing Interactive Systems (DIS 2017)*. Edinburgh, United Kingdom. 2017.

[13] **Anhong Guo**, Tim Paek. Exploring Tilt for No-Touch, Wrist-Only Interactions on Smartwatches. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services (MobileHCI 2016)*. Florence, Italy. 2016.
Best Paper Honorable Mention.

[14] **Anhong Guo**, Robert Xiao, Chris Harrison. CapAuth: Identifying and Differentiating User Handprints on Commodity Capacitive Touchscreens. In *Proceedings of the 10th ACM International Conference on Interactive Tabletops and Surfaces (ITS 2015)*. Madeira, Portugal. 2015.

[15] **Anhong Guo**, Shashank Raghu, Xuwen Xie, Saad Ismail, Xiaohui Luo, Joseph Simoneau, Scott Gilliland, Hannes Baumann, Caleb Southern, Thad Starner. A Comparison of Order Picking Assisted by Head-up Display (HUD), Cart-mounted Display (CMD), Light, and Paper Pick List. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers (ISWC 2014)*. Seattle, WA. 2014.
Best Paper Honorable Mention.