A Dynamic AI System for Extending the Capabilities of Blind People

Martin Grayson
Microsoft Research
Cambridge, UK
mgrayson@microsoft.com

Anja Thieme
Microsoft Research
Cambridge, UK
anthie@microsoft.com

Rita Marques
Microsoft Research
Cambridge, UK
t-rimarq@microsoft.com

Daniela Massiceti
Microsoft Research
Cambridge, UK
t-dmassi@microsoft.com

Ed Cutrell
Microsoft Research
Redmond, US
cutrell@microsoft.com

Cecily Morrison
Microsoft Research
Cambridge, UK
cecilym@microsoft.com

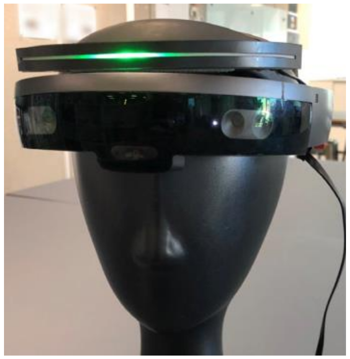


Figure 1: Modified HoloLens device with external LED display.

CHI 2020 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.
© 2020 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6819-3/20/04.
DOI: https://doi.org/10.1145/3334480.3383142

Abstract

We introduce an advanced computer vision-based AI system that offers people with vision impairments (VI) dynamic, in-situ access to information about the location, identity and gaze-direction of other people nearby. Our AI system utilizes the camera technology of a head-worn HoloLens device, which captures a near 180° field-of-view surrounding the person who is wearing it. Captured images are then processed by multiple state-of-the-art perception algorithms whose outputs are integrated into a real-time tracking model of all people that the system detected. Users can receive information about those people acoustically (via spatialized audio) using a wrist-worn input controller. Having such dynamic access to information, through AI system interactions, enables people with VI to: develop their communication skills; more easily focus on others; and be more confident in their social interactions. Thus, our work explores *how AI systems can serve as a useful resource for humans*, helping expand their agency to develop new or extend existing skills.

Author Keywords

AI system; design + AI; accessibility; blindness; visual impairment; disability; inclusive design; innovation.

CSS Concepts

• **Human-centered computing~Human computer interaction (HCI)**; *Accessibility*

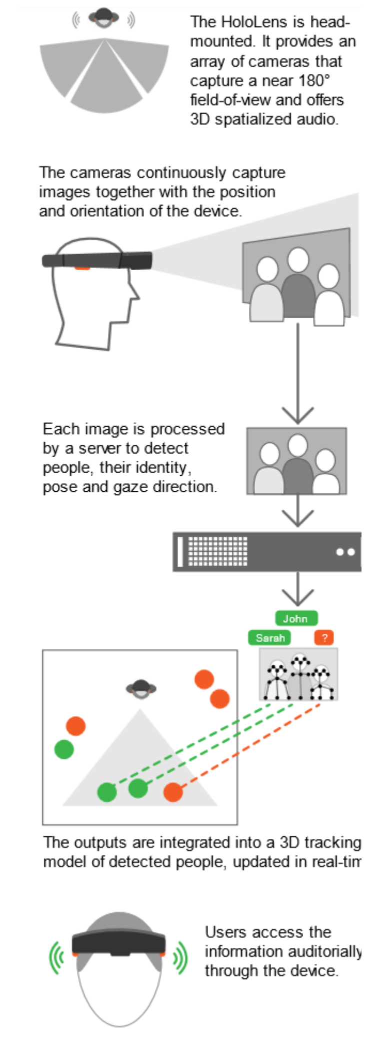

The HoloLens is head-mounted. It provides an array of cameras that capture a near 180° field-of-view and offers 3D spatialized audio.
The cameras continuously capture images together with the position and orientation of the device.
Each image is processed by a server to detect people, their identity, pose and gaze direction.
John
Sarah
?
The outputs are integrated into a 3D tracking model of detected people, updated in real-tim
Users access the information auditorially through the device.

Figure 2: Schematic representation of the AI system architecture and its workings.

Introduction

Advances in AI and machine learning enable the development of increasingly complex algorithmic models whose predictive capabilities invite exciting new possibilities for how these systems might support people [4]. In our research and AI system development with people with VI, we explore the design space for new kinds of human-AI partnerships and experiences that foreground human agency and sense-making. People with disabilities are often heavy technology users and have been early adopters of AI systems in their daily lives [1]. Through involving people with VI in multiple activities of design research [5] and ethnographic fieldwork [8], we learned that social relationships and interactions are critical for how blind people come to understand their surroundings, connect with others, or seek help. Based on this insight, we began to imagine how we could use AI technology to provide functionality that would offer people with VI dynamic, in-situ access to information about other people in the vicinity.

In our user research with blind adults and children, we started to see how such information access can create new opportunities for them to develop new or existing skills, and through this, to extend their capabilities. For example, for an adult, having a better understanding of who is nearby can make it easier for them to pro-actively approach someone to socialize rather than waiting for someone to reach out. It might also mitigate the potentially embarrassing situation of starting a conversation with someone who had quietly left the room. In our user research with a blind boy we further discovered the potential of using such an AI system to support the development of social communication skills, and an ability to focus on others, as failing to develop these fundamental social building blocks can have life-long consequences (*e.g.*, ⅔ of children born blind are being diagnosable with autism). As part of our demo, we will discuss our on-going user research.

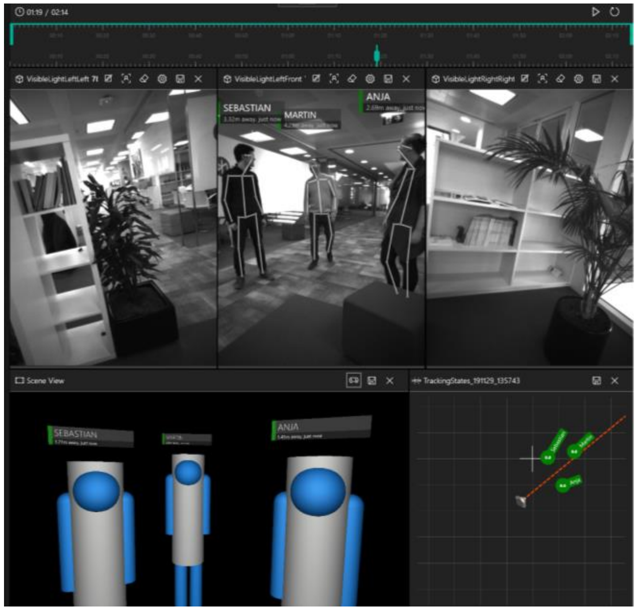


Figure 3. System view of the surroundings. Top: 180° camera field-of-view. Bottom: Integrated real-time 3D tracking model.

Overview of the AI System Architecture

The AI system that we developed comprises a head-worn HoloLens device, which we modified by removing the AR display lenses (Figure 1). The device captures a near 180° field-of-view surrounding the person who is wearing the device, tracks their head position, and provides high quality spatialized audio from non-occluding speakers above the ears (Figure 2).

Wrist-worn user input device

We adapted the O6 device (http://www.o6app.com) as our input device.

A single click activates the *overview* mode. In this mode, twist-interactions reveal additional information about each person.

A double click (de-)activates the *person-in-front* mode; and a triple click (de-)activates the *ambient* mode.

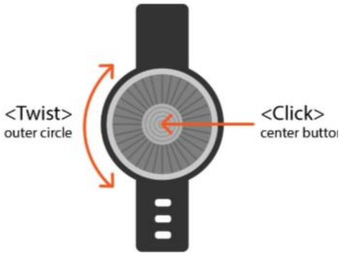


Figure 4: Look and configuration of the wrist-worn controller for accessing the experience modes.

A series of state-of-the-art computer vision algorithms process the captured images to continuously identify other people nearby. This includes algorithms that identify people [3], their location (re-implementation of [2]), activity, and gaze (implementation of [6]) in the vicinity. The outputs of these computer perception models are further merged into a maintained world state by a multi-target human tracker which runs about 12 to 15 frames a second (Figure 3). Compute is done on a box with two TitanX GPUs connected through WiFi to the HoloLens. Users receive information about people in the vicinity acoustically via spatialized audio. They can filter the information using a wrist-worn controller (Figure 4).

AI System Interactions & Experiences

We created three experience modes for users which provided easy access to information in different ways: *overview*, *person-in-front*, and *ambient* (Figure 5).

The *overview* mode was intended to help users build up a picture of who was around. For example, the user may find out that 'John' is in the room and can choose to approach him. In this mode, the system reads out the total number of people that it detects (e.g., "3 people"). Through twist-by-twist interactions, the user can receive additional details for each detected person: their name (or the system states "unknown"), approximate location, and time passed since the person was last detected, e.g., "John, near-front, 10 seconds ago". The user can either dwell to receive all those details, or quickly skip over these through further twists. Thus, this design approach enables users to quickly build up an understanding of their social surroundings, whilst giving them flexibility in how they engage with information, not forcing a set experience.

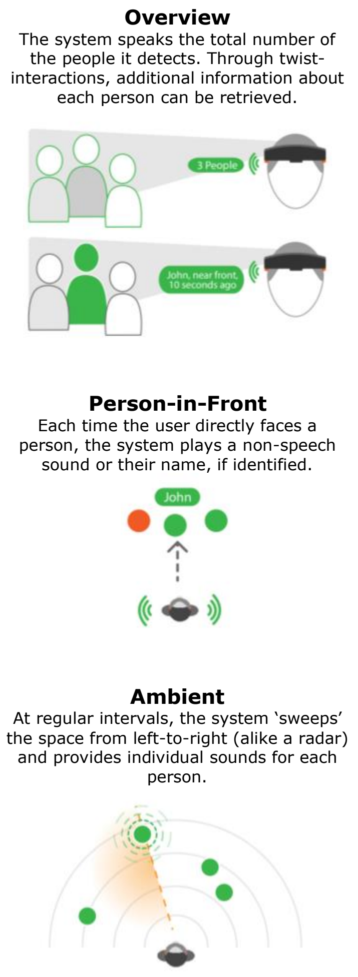
The *person-in-front* mode reads out a name whenever the user looks directly at another person. This might be used once the user is conversation with 'John' having spotted him using the *overview* mode. The users hears the name of the person, or if not identified, a spatialized sound to indicate the presence of a person. When the system has a robust internal representation of nearby people, this audio feedback is instant and presents as a short, comprehensive information cue. This functionality can help confirm who a conversation partner may be and enable the user to better adjust their body orientation towards that person.

As it can be difficult for people with VI to know how best to direct their head to help frame the faces of people for the AI system to recognize them [7], we developed *auditory orientation cues*. These changes in pitch if the user's head is tilted too high or too low, and 'snaps' to the nearest face. This enables the user to work with the system to support its ability to detect and identify people more robustly.

Finally, the *ambient* mode provides a sound for each person nearby at regular intervals (e.g., every 30 seconds) without any names or details. These interval-based updates allow users to have a more continuous sense of people's presence through peripheral, low-frequency audio that they can easily tune into, or ignore if irrelevant.

A Socially-Oriented AI System

Designed as a 'social' system that detects others nearby, we further considered how those bystanders would come to understand the purpose and functionality of the visible head-mounted device, as well as become an active participant in the social



Overview
The system speaks the total number of the people it detects. Through twist-interactions, additional information about each person can be retrieved.

Person-in-Front
Each time the user directly faces a person, the system plays a non-speech sound or their name, if identified.

Ambient
At regular intervals, the system 'sweeps' the space from left-to-right (alike a radar) and provides individual sounds for each person.

Figure 5: Schematic representations of the three core experience modes.

sense-making that the user is engaged in. To this end we affixed a semi-circular LED interface to the top of the HoloLens (Figure 1) that communicates the system state visually to bystanders. A moving 'white light' tracks the location of the nearest detected person and 'flashes green' when that person is identified by the system. The visual feedback enables bystanders to test-out the workings of the system and to use that understanding to also physically orientate themselves to the system to either make themselves more detectable to, or to hide from the system if they do not want to be captured. Creating transparency through a more open sharing of the system state may also help manage bystander expectations and ameliorate concerns that they otherwise might have about a system that would try to detect them 'unobtrusively'.

Conclusion

This will be the first time that our dynamic AI system is demonstrated at a scientific conference. Attendees will be able to wear the device and to explore for themselves the various auditory experience modes that we created using our wrist-controller. Furthermore, attendees will be able to inspect the underlying real-time 3D tracking model that we built. We will discuss our process and lessons learned from designing and prototyping a novel AI systems with people with VI as well as our concept and vision for creating a dynamic, AI-enabled resource for extending human capabilities.

Acknowledgements

We thank all participants who partnered with us in the system design and evaluation and the broader research team, who supported this project at different stages.

References

[1] Jeffrey P Bigham and Patrick Carrington. 2018. Learning from the Front: People with Disabilities as Early Adopters of AI. *HCIC*.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. *Proc. IEEE CVRP*, 7291–7299.

[3] Daniel Coelho de Castro and Sebastian Nowozin. 2018. From Face Recognition to Models of Identity: A Bayesian Approach to Learning about Unknown Identities from Unsupervised Data. *Proc. ECCV*, 745–761.

[4] Kori Inkpen, Stevie Chancellor, Munmun De Choudhury, Michael Veale, and Eric P. S. Baumer. 2019. Where is the Human?: Bridging the Gap Between AI and HCI. *Ext. Abstr. CHI 2019*. ACM, Paper W09, 9 pages.

[5] Cecily Morrison, Edward Cutrell, Anupama Dhareshwar, Kevin Doherty, Anja Thieme, and Alex Taylor. 2017. Imagining Artificial Intelligence Applications with People with Visual Disabilities using Tactile Ideation. *Proc. ASSETS 2017*. 81-90.

[6] Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. 2018. Deep directional statistics: Pose estimation with uncertainty quantification. *Proc. ECCV*, 534–551.

[7] Lee Stearns and Anja Thieme. 2018. Automated Person Detection in Dynamic Scenes to Assist People with Vision Impairments: An Initial Investigation. *Proc. ASSETS 2018*. 391-394.

[8] Anja Thieme, Cynthia L. Bennett, Cecily Morrison, Edward Cutrell, and Alex S. Taylor. 2018. "I can do everything but see!" -- How People with Vision Impairments Negotiate their Abilities in Social Contexts. *Proc. CHI 2018*. ACM, Paper 203.