Automated Person Detection in Dynamic Scenes to Assist People with Vision Impairments: An Initial Investigation

Lee Stearns[1] and Anja Thieme[2]
[1] University of Maryland, College Park, MD, USA; [2] Microsoft Research, Cambridge, UK
lstearns@umd.edu; anthie@microsoft.com

ABSTRACT
We propose a computer vision system that can automatically detect people in dynamic real-world scenes, enabling people with vision impairments to have more awareness of, and interactions with, other people in their surroundings. As an initial step, we investigate the feasibility of four camera systems that vary in their placement, field-of-view, and image distortion for: (i) capturing people generally; and (ii) detecting people via a specific person-pose estimator. Based on our findings, we discuss future opportunities and challenges for detecting people in dynamic scenes, and for communicating that information to visually impaired users.

Author Keywords
Accessibility; vision impairment; context-aware computing; computer vision; wearables; person detection; tracking.

ACM Classification Keywords
Human-centered computing → Accessibility technologies

INTRODUCTION
Recent research into the design of computer vision (CV) systems to assist people with visual impairments (VI) has started to move beyond supporting independent activities of daily living (*e.g.*, recognizing text [13] or objects [19], and aiding navigation [8]), and toward a closer consideration of how people with VI are inter-connected with, and supported by, others [5, 6, 14, 23, 24, 26, 30]. Examples include crowd-sourced answers to visual questions [4, 9] and CV-assisted social experiences—for example making the capture [18, 25] or editing and sharing of photos [3] more appealing to people with VI. Wu *et al.* [27] used computer vision to automatically integrate accessible alt-text information with Facebook photos, allowing blind users to feel more included and engaged with conversation around photos. Moving from online photo consumption to real-world situations, Zhao *et al.* [28] developed a Facebook Messenger bot that processes images from a smartphone camera in real-time to provide users with information about the number of people in front of them, and offers additional information such as the identity, relative location, and facial features of Facebook friends via a screen reader.

Building on this research, we investigate the use of CV to detect people in dynamic, real-world situations. Our aim is to help people with VI gain awareness of the people in their immediate surroundings to facilitate social interactions [10, 24, 28] and to protect their privacy and personal safety [1, 7]. Advancing research by Zhao *et al.* [28], we seek to develop a model of peoples' identities, locations, and movements in 3D spaces based on *continuous* processing and recognition using body-worn or stationary cameras. Building a *live* model of other people nearby may mitigate the challenge of requiring users to frame people in front of a camera, and it enables technology to *automatically* provide important context information to a user that otherwise could be missed.

While previous work in CV has explored challenges such as how to integrate different vision modules to reliably detect people [15], or how to track the behaviour of crowds [21] in the real world, we wanted to better understand—as a first step—how the choice of camera can impact the accuracy of people detection. Comparing four different camera systems, we derive: (i) insights into their feasibility for detecting people; and (ii) key challenges for intersecting CV and interaction design to help guide future research in this space.

CAPTURING PEOPLE: A CAMERA COMPARISON PILOT
In our feasibility pilot, we compared the performance of two *head-mounted wearable cameras*, a HoloLens[i] and a Rico Theta[ii], and two *stationary cameras*, an iPhone with a 238˚ wide-angle lens and a 360˚ Polycom[iii] conferencing system (Figure 1 left). These cameras further varied in their field-of-view (from narrow 45˚ to spherical 360˚) and imaging capabilities (frame size, resolution, distortion; Figure 1 center), thus representing a breadth of possible camera systems. To test their performance, we set up a small, semi-controlled meeting scenario (Figure 1 right), that involved five participants (P1–P5) who were instructed to: (1) arrive individually, (2) take a seat at the table, talk to each other for a few minutes, and (3) then to depart in pairs. An additional person, the *user*, was wearing the HoloLens and Ricoh Theta on their head along with a blindfold to simulate the experience of restricted sight. This method can be useful in early usability-focused (rather than empathy/ability-focused) system testing [17, 20], but we acknowledge that the head movements of a blindfolded user can be quite different from those of a blind person. We simultaneously recorded video with each of the four cameras and an overhead camera to establish a ground truth for

ASSETS '18, October 22–24, 2018, Galway, Ireland

ACM ISBN 978-1-4503-5650-3/18/10.
https://doi.org/10.1145/3234695.3241017

Figure 1. ***Left:*** **Camera placement and field-of-view (FoV);** ***Center*****: Sample images from each camera;** ***Right*****: Test scenario set-up.**

each person's position. While our test scenario was only brief (8 mins, 14 secs), the resulting video stream of the five cameras consisted of over 65,000 individual frames. To efficiently and reliably annotate these frames, we implemented Cao *et al.*'s [12] state-of-the-art pose-detection algorithm to automatically recognize multiple people in RGB images, corrected any false outputs, and added additional labels manually via a simple custom-built interface.

FINDINGS

Unsurprisingly, the two 360˚ cameras performed best in terms of ensuring each participant was visible at all times, while the iPhone with wide-angled lens captured the majority of the room except for one participant seated behind it (P5). The HoloLens, with its narrow 45˚ field-of-view (FoV), only captured the two participants who sat in front of the user (P2 and P3) with any regularity, and at best for 75% of the frames. This confirms the importance of a camera's field-of-view and position for capturing people in a real-world scene.

Next, we assessed the person detection performance for the frames where a person was visible. The detection algorithm achieved a high average accuracy for most cameras (up to 96.9%), aside from the HoloLens (avg. 56.8%). In 31.2% of the HoloLens frames, the user's head was tilted too low, cutting off participants' heads. Interestingly, the detection algorithm still functioned reasonably well for detecting torsos (up to 73.5%), even when a person's head and legs were occluded. Further, false positive rates for the automated detection algorithm were very low across all cameras ($\leq 2\%$), and almost non-existent for the Ricoh Theta (~0.1%). This result was surprising considering the radial and fisheye lens distortions of the wide-angle and 360˚ cameras (Figure 1). This suggests that camera systems with such distortions do not need to be excluded from automatic image processing but can be beneficial for capturing people in a dynamic scene due to their wider FoV.

DISCUSSION & CHALLENGES FOR FUTURE WORK

While limited in scope, and excluding important factors such as the distance, density, and movements of people, our pilot shows the effects of camera positioning, field-of-view, and distortion for detecting people in a dynamic scene. We close with a discussion of implications for future vision systems designed to increase awareness of others for people with VI.

Beyond Capture: Anchors to Assist in Data Interpretation

In our pilot, the 360˚ Ricoh Theta performed best at capturing and detecting people, avoiding challenges with aiming the camera, and, as a wearable, providing an ego-centric rather than stationary frame of reference. Beyond capture, we will need to consider how to best convey data about the relative positions and movements of other people to the user. In Zhao *et al.* [28], the phone acted as a 'frame of reference'. Identifying a focus to spatially anchor available information about people—their positions in relation to the user and each other—is more complicated for a live 360˚ system, however. Here, a body-worn camera, especially if head-mounted, has an advantage, allowing the system's focus to be aligned with the user's head or gaze-direction to help anchor information in a dynamic scene. However, a head-worn design may also present barriers to social acceptance [29], suggesting more subtle designs would be preferable (*i.e.*, [16]). Yet, if the design is too subtle, others may not recognize the camera, impacting potential choices to opt out, or protect privacy.

From Calibration to Collaboration with a Vision System

Most detection errors in the HoloLens data related to the orientation of the user's head. Camera calibration challenges are addressed in designs for blind photography [2, 18] and other camera-related detection tasks [22] that often provide audio cues or haptic vibrations to assist in image framing and to prevent blur. Our pose-detection algorithm performed well in detecting people despite occluded body parts (*e.g.*, torso detected, but no head), suggesting that this information could be used to signal users to look up to help improve detection accuracy. Constant calibration and collaboration with a system are more complicated in dynamic scenes, where people move continuously, requiring efficient and intuitive interactions that do not become burdensome or distracting.

Beyond a Vision Snapshot: Accounting for Temporality

All context-aware systems [11] based on sensing should account for changes due to motion as well as technology errors in detections (*e.g.*, algorithm uncertainty, restricted field-of-view and occlusions). We believe that by developing a continuous model of nearby people that tracks each person and their location, we can achieve more robust in-situ person detection. Further, temporal information about when a person was last 'seen' by the system (*e.g.*, just now *vs.* 15 secs ago), could prompt the user to either look around to assist in a re-detection of the person, or to be more cautious in interpretations of system feedback (*e.g.*, the person may have left the scene). Future work will need to explore how factors such as temporality can best be communicated to aid a user's understanding of a social scene in meaningful ways.

ACKNOWLEDGEMENTS

We thank our team, especially: Ed Cutrell, Cecily Morrison, Martin Grayson, Stephan Garbin, for aiding this research.

REFERENCES

1. Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. 2015. Privacy Concerns and Behaviors of People with Visual Impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, 3523-3532. https://doi.org/10.1145/2702123.2702334
2. Jan Balata, Zdenek Mikovec, and Lukas Neoproud. 2015. BlindCamera: Central and Golden-ratio Composition for Blind Photographers. In *Proceedings of the Mulitimedia, Interaction, Design and Innnovation* (MIDI '15). ACM, Article 8, 8 pages. Dhttps://doi.org/10.1145/2814464.2814472
3. Cynthia L. Bennett, Jane E, Martez E. Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18). ACM, Paper 76, 12 pages. https://doi.org/10.1145/3173574.3173650
4. Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (UIST '10). ACM, 333-342. https://doi.org/10.1145/1866029.1866080
5. Stacy M. Branham and Shaun K. Kane. 2015. Collaborative Accessibility: How Blind and Sighted Companions Co-Create Accessible Home Spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 2373-2382. http://dx.doi.org/10.1145/2702123.2702511
6. Stacy M. Branham and Shaun K. Kane. 2015. The Invisible Work of Accessibility: How Blind Employees Manage Accessibility in Mixed-Ability Workplaces. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (ASSETS '15), 163-171. http://dx.doi.org/10.1145/2700648.2809864
7. Stacy M. Branham, et al. 2017. "Is Someone There? Do They Have a Gun": How Visual Information about Others Can Improve Personal Safety Management for Blind Individuals. *Proc. ASSETS 2017*. ACM, 260-269. https://doi.org/10.1145/3132525.3132534
8. J.M. Hans du Buf, João Barroso, João M.F. Rodrigues, Hugo Paredes, Miguel Farrajota, Hugo Fernandes, João José, Victor Teixeira, and Mário Saleiro. 2011. The SmartVision Navigation Prototype for Blind Users. *International Journal of Digital Content Technology and its Applications 5*(5), 351-361. http://hdl.handle.net/10400.1/893
9. Michele A. Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P. Bigham, and Amy Hurst. 2012. Crowdsourcing subjective fashion advice using VizWiz: challenges and opportunities. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility (ASSETS '12).* ACM, 135-142. http://dx.doi.org/10.1145/2384916.2384941
10. Verena R. Cimarolli, Kathrin Boerner, Mark Brennan-Ing, Joann P. Reinhardt, and Amy Horowitz 2012. Challenges faced by older adults with vision loss: a qualitative study with implications for rehabilitation. *Clinical rehabilitation*, *26*(8), pp.748-757. http://journals.sagepub.com/doi/abs/10.1177/0269215511429162
11. Anind K. Dey. Context-Aware Computing. In *Ubiquitous Computing Fundamentals 2016*, Chapman and Hall/CRC, pp. 335-366.
12. Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2016. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050.* https://arxiv.org/abs/1611.08050
13. Xiangrong Chen and Alan L. Yuille. 2004. Detecting and reading text in natural scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* vol. 2. http://dx.doi.org/10.1109/CVPR.2004.1315187
14. William Easley, Michele A. Williams, Ali Abdolrahmani, Caroline Galbraith, Stacy M. Branham, Amy Hurst, and Shaun K. Kane. 2016. Let's Get Lost: Exploring Social Norms in Predominately Blind Environments. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '16). ACM, 2034-2040. https://doi.org/10.1145/2851581.2892470
15. Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, IEEE, 1-8. https://doi.org/10.1109/CVPR.2008.4587581
16. FITT360. The First 360° Neckband Wearable Camera. Unobtrusive and effortless way to capture and share your moments in 360° true First-Person-View. Last retrieved 15th of June 2018, from https://www.kickstarter.com/projects/467094941/fitt360-the-first-360-neckband-wearable-camera
17. Luis A. Guerrero, Francisco Vasquez, and Sergio F. Ochoa. 2012. An indoor navigation system for the visually impaired. *Sensors* 12, no. 6 (2012): 8236-8258. http://dx.doi.org/10.3390/s120608236

18. Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. 2011. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility* (ASSETS '11). ACM, 203-210. http://dx.doi.org/10.1145/2049536.2049573

19. Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. 2017. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17).* ACM, 5839-5849. https://doi.org/10.1145/3025453.3025899

20. David McGookin, Stephen Brewster, and WeiWei Jiang. 2008. Investigating touchscreen accessibility for people with visual impairments. *Proc. NordiCHI 2008*. ACM, 298-307. http://dx.doi.org/10.1145/1463160.1463193

21. Mikel Rodriguez, Josef Sivic, and Ivan Laptev. 2017. The Analysis of High Density Crowds in Videos. In *Group and Crowd Behavior for Computer Vision*, 2017, 89-113. https://doi.org/10.1016/B978-0-12-809276-7.00006-0

22. Seeing AI. *Talking Camera for the Blind.* Last retrieved 15th June 2018, from https://www.microsoft.com/en-us/seeing-ai/

23. Sarit Felicia Anais Szpiro, Shafeka Hashash, Yuhang Zhao, and Shiri Azenkot. 2016. How People with Low Vision Access Computing Devices: Understanding Challenges and Opportunities. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '16), 171-180. https://doi.org/10.1145/2982142.2982168

24. 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. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, Paper 203, 14 pages. https://doi.org/10.1145/3173574.3173777

25. Marynel Vázquez and Aaron Steinfeld. 2012. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility* (ASSETS '12). ACM, 95-102. http://dx.doi.org/10.1145/2384916.2384934

26. Michele A. Williams, Amy Hurst, and Shaun K. Kane. 2013. "Pray before you step out": describing personal and situational blind navigation behaviors. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '13). ACM, Article 28, 8 pages. http://dx.doi.org/10.1145/2513383.2513449

27. Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17).* ACM, 1180-1192. https://doi.org/10.1145/2998181.2998364

28. Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2018. A Face Recognition Application for People with Visual Impairments: Understanding Use Beyond the Lab. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, Paper 215, 14 pages. https://doi.org/10.1145/3173574.3173789

29. Kristen Shinohara and Jacob O. Wobbrock. 2016. Self-Conscious or Self-Confident? A Diary Study Conceptualizing the Social Accessibility of Assistive Technology. *ACM Trans. Access. Comput. 8, 2, Article 5*, 31 pages. http://dx.doi.org/10.1145/2827857

30. Annuska Zolyomi, Anushree Shukla, and Jaime Snyder. 2016. Social Dimensions of Technology-Mediated Sight. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '16). ACM, 299-300. https://doi.org/10.1145/2982142.2982190

---

[i] https://www.microsoft.com/en-us/hololens

[ii] https://theta360.com/uk/

[iii] http://www.polycom.co.uk/