ABSTRACT

Title of Dissertation: EGOCENTRIC VISION IN ASSISTIVE TECHNOLOGIES FOR AND BY THE BLIND

Kyungjun Lee, Doctor of Philosophy, 2022

Dissertation Directed by: Hernisa Kacorri
College of Information Studies & Department of Computer Science

Visual information in our surroundings, such as everyday objects and passersby, is often inaccessible to people who are blind. Cameras that leverage egocentric vision, in an attempt to approximate the visual field of the camera wearer, hold great promise for making the visual world more accessible for this population. Typically, such applications rely on pre-trained computer vision models and thus are limited. Moreover, as with any AI system that augments sensory abilities, conversations around ethical implications and privacy concerns lie at the core of their design and regulation. However, early efforts tend to decouple perspectives, considering only either those of the blind users or potential bystanders.

In this dissertation, we revisit egocentric vision for the blind. Through a holistic approach, we examine the following dimensions: type of application (objects and passersby), camera form factor (handheld and wearable), user's role (a passive consumer and an active director of technology), and privacy concerns (from both end-users and bystanders). Specifically, we propose

to design egocentric vision models that capture blind users' intent and are fine-tuned by the user in the context of object recognition. We seek to explore societal issues that AI-powered cameras may lead to, considering perspectives from both blind users and nearby people whose faces or objects might be captured by the cameras. Last, we investigate interactions and perceptions across different camera form factors to reveal design implications for future work.

Egocentric Vision in Assistive Technologies For and By the Blind

by

Kyungjun Lee

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Hernisa Kacorri, Chair
Professor Michelle Mazurek
Professor Abhinav Shrivastava
Professor Gregg Vanderheiden
Professor Hal Daumé III

© Copyright by
Kyungjun Lee
2022

Dedication

To my parents, Kiyoung Kim and Kichoon Lee, and my life partner, Nayoon Lee, who have always been with me during this journey.

Acknowledgments

I would like to thank my advisor Hernisa Kacorri. She introduced me to the accessibility research field and has supported all of the research work presented in this dissertation. Her passionate advice has motivated me to keep being curious, asking questions, and seeking answers to those questions. As she has done, I will strive for becoming a positive influence on others. Furthermore, I am grateful to Abhinav Shrivastava for providing insightful guidance, which has helped me to advance my research in a multidisciplinary area, and Hal Daumé III for his enthusiasm for just technologies, which has influenced not only my research but my life motto. I would like to express my appreciation to Michelle Mazurek and Gregg Vanderheiden, who have spared their priceless time sharing constructive feedback on my dissertation.

My dissertation has also benefited from the influence and support of my colleagues. Many thanks to Chieko Asakawa and Daisuke Sato for their guidance and support. Their research effort has encouraged me to consider the broader implications of my research. I would like to thank my friends and labmates whose presence has been of vital importance during many years at UMD. Special thanks to Jonggi Hong for leading me to Hernisa's lab and Ebrima Jarjue for discussing ideas together. Thanks also to the IAM and HCIL labmates for the encouragement and lively conversations on the second floor of the Hornbake building: Utkarsh Dwivedi, Jaina Gandhi, Mayanka Jha, June Xu, Raj Parikh, Sravya Amancherla, Simone Pimento, Tzu-Chia Yeh, Rie Kamikubo, Julia Kerns, Elizabeth Beers, Ernest Essuah Mensah, Lining Wang, Seokbin Kang,

Teja Maddali, Pramod Chundury, Sigfried Gold, Emma Dixon, Sungbok Shin, Alisha Pradhan, Biswaksen Patnaik, Xin Qian, Salma Elsayed-Ali, Yuhan Luo, Young-Ho Kim, Yimin Xiao, and Ruipu Hu. Thanks additionally to Saki Asakawa, Eagle Dapeng Zhao, and Emily Pan for helping conduct user studies and being great lunchmates during my time at CMU. Also, my research would have not been possible without all participants in our user studies. I appreciate all of them sharing their lived experiences.

Last but not least, I am sincerely grateful to my life partner, Nayoon Lee, and my family, Kiyoung Kim, Kichoon Lee, and Kyunghwa Lee, for their endless support and love—thank you for everything.

Table of Contents

Dedication ii

Acknowledgements iii

Table of Contents v

List of Tables vii

List of Figures ix

Chapter 1: Introduction 1
- 1.1 'Application' Dimension . . . 3
- 1.2 'Camera Form Factor' Dimension . . . 5
- 1.3 'User Role' Dimension . . . 6
- 1.4 'Privacy Concern' Dimension . . . 7
- 1.5 Overview of This Dissertation . . . 7

Chapter 2: Related Work 10
- 2.1 Computer Vision for Blind People . . . 11
- 2.2 Blind People's Camera Interactions . . . 17

Chapter 3: Hands Holding Clues in Teachable Object Recognition on Hand-Held and Head-Worn Cameras 21
- 3.1 Motivation . . . 22
- 3.2 Methods . . . 24
- 3.3 Results . . . 33
- 3.4 Discussion . . . 41
- 3.5 Summary . . . 45

Chapter 4: Hand-Guided Sonification for Teachable Object Recognition on Hand-Held Camera 46
- 4.1 Motivation . . . 48
- 4.2 Methods . . . 52
- 4.3 Results . . . 61
- 4.4 Discussion . . . 72
- 4.5 Summary . . . 75

Chapter 5: Hand-Primed Object Localization and Recognition on Hand-Held and Head-Worn Cameras 76
5.1 Motivation . . . 79
5.2 Methods . . . 81
5.3 Results . . . 90
5.4 Discussion . . . 102
5.5 Summary . . . 105

Chapter 6: Hand-Primed Sonification for Teachable Object Recognition on Head-Worn Camera 106
6.1 Motivation . . . 107
6.2 Methods . . . 109
6.3 Results . . . 115
6.4 Discussion . . . 125
6.5 Summary . . . 127

Chapter 7: Blind People's Interactions with Pedestrian Detection System on Head-Worn Camera 129
7.1 Motivation . . . 132
7.2 Methods . . . 133
7.3 Results . . . 140
7.4 Discussion . . . 152
7.5 Summary . . . 157

Chapter 8: Social Acceptance of Pedestrian Detection System on Head-Worn Camera 159
8.1 Motivation . . . 162
8.2 Methods . . . 163
8.3 Results . . . 174
8.4 Discussion . . . 186
8.5 Summary . . . 188

Chapter 9: Conclusion 189
9.1 Contributions . . . 189
9.2 Future Directions . . . 192
9.3 Epilogue . . . 194

Bibliography 196

List of Tables

2.1 Input, model, and prediction characteristics in prior object recognition for the blind. 12

3.1 Overview of the dataset characteristics used in our work. An asterisk indicates a simulated environment. 27

4.1 Comparison of several prior applications addressing camera manipulation for people with visual impairments. 49
4.2 Participants' demographics and years of smartphone use. Light perception is indicated with an asterisk. 56
4.3 The number of participants who included variations while taking photos in *train*, *vanilla test*, and *wild test* 65

5.1 Quantitative analysis of the hand segmentation models. The mean intersection of union ($mIoU$) and the COCO standard metrics (AP, AP_{50}, AP_{75}) are used. . . . 91
5.2 Quantitative analysis of the object localization models. The mean intersection of union ($mIoU$) and the COCO standard metrics (AP, AP_{50}, AP_{75}) are used. . . . 94
5.3 Object localization performances of the methods on a different set of the TEgO data. For this analysis, the data from the blind person are separated from the data from the sighted person. The bold indicates the best among the models except for *Faster R-CNN (w/ whole bbox)*. 98
5.4 Quantitative analysis of object classification on the TEgO dataset. F1-score, precision, and recall (*macro average*) are used. 99
5.5 Object classification performances on the blind individual's data in TEgO. The methods were trained and tested on TEgO (blind). 101

6.1 Participants' demographics and prior knowledge with machine learning. Light perception is indicated with an asterisk. Some demographic information for P2 and P9 are not presented here, which is marked with dash (-), due to their request. 115

7.1 Demographic information of our participants. Participants with light perception are marked with an asterisk in *Vision level*. 137
7.2 Attributes in our annotation. 139
7.3 Performance of our prototype system on the camera frames from blind participants. 146
7.4 Information about non-handheld cameras designed for helping users with visual impairments access visual surroundings. 153

8.1 Video characteristics and reference codes. 164
8.2 Demographic information about blind participants. 171

8.3 Categories and descriptions of social acceptance questions for blind participants, and mean (std.) of scores of their responses to the questions. Responses are converted to numeric values (-3 to 3): -3=*Strongly disagree*, 3=*Strongly agree*. . 173

List of Figures

1.1 This dissertation considers the following dimensions in the context of assistive technologies employing egocentric vision with blind users: (i) the application and whether it focuses on objects or passersby, (ii) the camera form factor and whether it is hand-held or head-worn, (iii) the role of the user in the machine learning model and whether they are considered as active directors versus passive consumers, and (iv) whether privacy concerns are explored or not. 2

3.1 An illustration of our hand-guided object recognition approach on an example from our egocentric dataset. Given a photo of an object in proximity to a hand, it first identifies the hand and then estimates the object center, which is then cropped and passed to the recognition model. 22

3.2 In our approach, a hand segmentation model (Step I) is fine-tuned to estimate the center of the object in proximity to the hand (Step II). A bounding box, placed in that center is used to isolate the object and crop the image, which is then passed to the object classification model (Step III). 25

3.3 Examples of input, annotation, and output for our hand segmentation and object localization models. 27

3.4 Examples from each dataset. Glassense-Vision, VizWiz, and our benchmark examples are selected to include hands. 28

3.5 Nineteen objects used in our data collection. Objects in the same category are displayed in proximity. 31

3.6 Positive and negative outputs of our object localization model on the Glassense-Vision and VizWiz datasets. 36

3.7 Our hand-guided object recognition method (CO) tends to improve recognition accuracy on average for S and B compared to the original HO and O methods. . . 38

3.8 Accuracy gain of our method (CO) over HO and O is more pertinent in cluttered backgrounds (wild). 39

3.9 On average CO outperforms HO and O consistently across training sample sizes $k = 1, 5, 20$. 39

3.10 Presence of hands (HO and CO) seems to have a different effect for generic vs. teachable models. 39

3.11 Positive and negative results on TEgO, with out-of-frame hands for some of the negative examples. 41

3.12 Confusion matrix for the CO models showing that misclassification occurs within objects of similar shape. Cans and bottles are indicated as “-c” and “-b”, respectively. 42

4.1 Examples of challenging photos taken by people with visual impairments sent to Vizwiz crowdworkers. About 28% of all photos are categorized as unanswearable [1]. 47
4.2 Training examples in our hand-segmentation and object-localization models using the EgoHands, GTEA, GTEA Gaze+, Intel Egocentric, and TEgO datasets (in the left-to-right order). Original images are shown on the first row, hand segmentation on the second row, and object center annotations on the last row. EgoHands is used only in hand segmentation. 51
4.3 Our nonvisual feedback mechanism. 52
4.4 The architecture of our real-time feedback approach for better camera framing. Our testbed sends an image to the server, receives an estimated location, and triggers the corresponding feedback. 52
4.5 Technology experience and attitude responses. All participants have smartphones; more than half are using apps for object recognition; and all are positive about technology. 56
4.6 Objects for (a) practice and (b) stimuli in our study. 58
4.7 The wild test setting in the cluttered environment. 59
4.8 Vanilla environment and the camera manipulation across participants in train mode. 60
4.9 Average model accuracy per participant. 62
4.10 Proportion of photos with fully, partially, or not included objects and hands. Errors bars show variance among participants, who were able to fully capture the object in 60–80% of their photos and include their hand in more than 50%. 64
4.11 Post-study feedback. Most participants thought training is not difficult and they were willing to provide photos to train and improve the performance of an object recognizer. 65
4.12 By representing the stream of interactions in a (participant, object) pair as a unique node, we find three distinct clusters and their descriptive features that are most responsible for this cluster formation. 66
4.13 Comparing clusters across different interaction characteristics during training. . . 69
4.14 Cluster distribution across participant-object pairs. 69

5.1 Examples of egocentric photos taken by people with visual impairments for object recognition in the crowdsourcing app, VizWiz [1], and a personalized fine-grained object recognition model in TEgO [2]. These examples illustrate the need for object-of-interest localization for better camera framing. 77
5.2 The architecture of our hand-primed object recognition model. It consists of two neural networks: one (above) for hand segmentation and another (below) for object localization and recognition. The hand segmentation network is pre-trained, and its output is infused into the other model to help recognize the object of interest being interacted by the egocentric hand. A fully-connected layer for classification is appended to *conv7* of the object localization model. It allows object classification to share the *conv1–conv7* layers of object localization. 80
5.3 Input data for training models: an original image, its hand mask, and its object center annotation. 83

5.4 Input data for two different versions of Faster R-CNN: an original image, a whole object bounding box, and an object center bounding box. 89
5.5 Hand segmentation outputs of the comparison models and our approach. The hand segmentation is overlaid with the green color. All the models except for the multi-class models appropriately segmented whole hand(s) from the testing examples; the multi-class models failed to segment fingers from the examples from GTEA and GTEA Gaze+. 93
5.6 Object localization outputs of the comparison models our approach. The object localization output is overlaid with the blue color on the testing examples; the threshold for the localization output was set to 0.5. In comparisons with *No-HandData*, our method (*Ours*) shows that the hand information helps localize the target object. 95
5.7 Failure cases of our method despite the appropriate hand segmentation outputs on these testing examples. 96
5.8 Cases when the classification task helps object localization. The previous models without classification fail to locate the target object, but become able to localize it after being trained for classification. 100
5.9 Examples show that models sometimes become confused or unable to locate an object of interest when trained for both object localization and classification (*With Classification*). 101

6.1 System architecture of our sonification feedback designed for smartglasses for the blind. The overall diagram is similar to that in Chapter 4. The main difference is that the smartglasses one is powered by our hand-primed object localization model introduced in Chapter 5 and does not provide the haptic feedback when the object is at the center of camera frame. 109
6.2 Our sonification feedback mechanism designed for smartglasses for blind users. . 110
6.3 Average model accuracy per participant. 116
6.4 By representing the stream of interactions in a (participant, object) pair as a unique node, we find three distinct clusters and their descriptive features that are most responsible for this cluster formation. 121
6.5 Cluster distribution across participant-object pairs. 122
6.6 Comparing clusters across different interaction characteristics during teaching. . . 123

7.1 A pedestrian detection scenario demonstrating the data collection in our study. A blind person, wearing smartglasses with our working prototype, and a sighted person walk toward each other in a corridor. The smartglasses detect the sighted passerby, estimate his proximity, and share his relative location and head pose. . . 132
7.2 System diagram describing the procedure of pedestrian detection: (1) The client app on smartglasses captures an image and sends it to the remote server; (2) Once the remote server detects a passerby, it estimates the distance, position, and head pose of the passerby; (3) The output is then sent back to the smartglasses; (4) The smartglasses deliver the output to the blind user via audio. 134

7.3 Diagram of study procedure. A blind participant walked eight times in a corridor that is 66 feet (20 meters) long and 79 inches (2 meters) wide. Each time a different sighted person walked towards them. In four of the eight times, those individuals were sighted study participants. In the other four, they were sighted members in our research lab. In addition to the proxemic signals, GlAccess can recognize the lab members, simulating future teachable applications [3] where users can personalize the model for familiar faces. 136
7.4 Diagrams presenting inclusion rates of any body part or head of passersby in camera frames per distance. A relatively high percentage of camera frames capture a passerby (face and body) at a distance larger than 6 feet in a corridor that is 79 inches wide; this can vary across blind participants with some outlier cases being as low as 50%. Within 6 feet it can get unpredictable with faces often being included less. 142
7.5 Heatmap diagram presenting the rates of body inclusion patterns in frames across distances, interactions, and participants. 144

8.1 Video stimuli shown to our sighted participants during the online and in-situ studies. Both perspectives were included in the online study; in-situ only the user perspective. 160
8.2 Diagram of data processing of *GlAccess*. 169
8.3 Starting points and walking directions of blind participants and sighted passerby participants in a 66-feet (20-meter) corridor. A blind participant wearing *GlAccess* and a sighted passerby participant walk toward each other. 173
8.4 Online sighted participants' opinions on use cases of assistive wearable cameras. Participants were asked whether they would be okay with each of the use cases. . 175
8.5 Comparisons of online participants' responses. 176
8.6 Online sighted participants' responses to the social acceptance questions before and after the user perspective video. Responses are converted to numeric values (-3 to 3): -3=*Strongly disagree*, 0=*Neutral*, 3=*Strongly agree*. 178
8.7 The four figures on the left for comparison of responses from in-situ sighted participants with those from online sighted participants, and two figures on the right for comparison of in-situ sighted participants' responses before and after the user video. 181
8.8 In-situ sighted participants' opinions on use cases of assistive wearable cameras. Participants were asked whether they would be okay with each of the use cases. . 182
8.9 In-situ sighted participants' responses to the social acceptance questions before and after the user video. Responses are converted to numeric values (-3 to 3): -3=*Strongly disagree*, 0=*Neutral*, 3=*Strongly agree*. 183
8.10 Participants' opinions on feature detection. 184

Chapter 1: Introduction

Visual information in our surrounding is inaccessible to blind people. Many objects are not recognizable without sight unless one learns to associate and remembers this association of objects with information available via the other senses (*i.e.*, sound, smell, taste, and touch) or their location (*e.g.*, some ad-hoc organization scheme). Similarly, visual cues such as eye gaze and distance from people in one's surroundings such as passerby pedestrians are typically inaccessible to blind people. This is critical not only for fostering social interactions but also for ensuring one's safety *e.g.* in practicing social distancing as this recent pandemic taught us.

To support blind people in accessing their visual surroundings, researchers, designers, and end users have been exploring assistive technologies that employ cameras and egocentric vision; in this dissertation, with egocentric vision we refer to computer vision efforts that aim to understand visual data such as images and videos, captured from the user's perspective. In particular, assistive technologies that leverage egocentric vision typically aim to convert visual data into other sensory data, such as auditory and haptic feedback, which can be more accessible to blind users.

Yet, the underlying data-fueled algorithms in these assistive technologies, typically rely on machine learning models that are pre-trained on data sourced from sighted people and tasks that are typically defined in broader computer vision tasks (*e.g.*, meant to navigate a robot). These

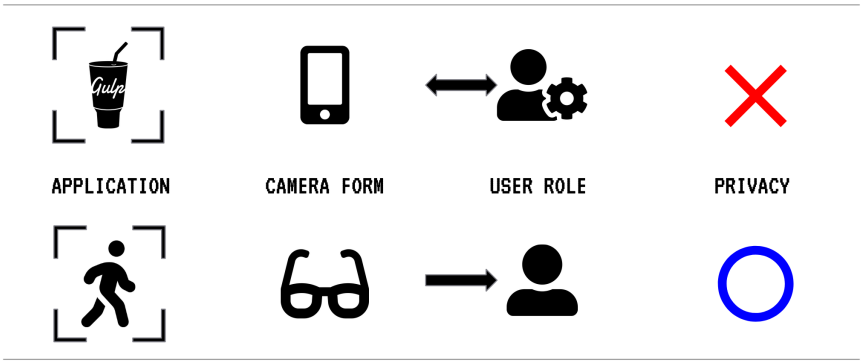


Figure 1.1: This dissertation considers the following dimensions in the context of assistive technologies employing egocentric vision with blind users: (i) the application and whether it focuses on objects or passersby, (ii) the camera form factor and whether it is hand-held or head-worn, (iii) the role of the user in the machine learning model and whether they are considered as active directors versus passive consumers, and (iv) whether privacy concerns are explored or not.

limitations may restrict the utility of computer vision in assistive technologies for the blind. For example, a blind user may be able to tell just by touch whether the object in front of them is a soda bottle (typical object recognition) but may not be able to distinguish whether it is a diet or regular coke (*i.e.*, instance-level object recognition); even this simple distinction can be critical for some (*e.g.*, those who are blind and diabetic). In general, building a super object classifier that can recognize all possible object instances for all blind people is not feasible with current object recognition technology [4]. Furthermore, studying social acceptance of such assistive technologies given the always-on nature of their cameras has become imperative, as it can impact both their design and regulation. Early efforts aiming to understand these ethical implications and privacy concerns tend to be one-sided; consider either blind users or potential bystanders without coupling perspectives from both of these stakeholders in their studies.

These considerations shape this dissertation on AI-infused assistive cameras for blind people, where a holistic approach is taken to explore egocentric vision for assistive technologies for and by the blind. As shown in Figure 1.1, this dissertation considers four dimensions with

their motivation discussed more in depth below (Section 1.1-1.4). The first is the context of the ***application***, where we examine egocentric vision for object and pedestrian detection. The second aspect is the ***camera form factor*** since photos and videos taken by blind people can exhibit unique characteristics and are affected by the camera manipulation. In this dissertation, we focus on hand-held and head-worn cameras typically deployed in smartphones and smart glasses, respectively. Our third and most important consideration is the ***user role*** in the context of the computer vision models. We consider users in a passive-consumer role when the machine learning models deployed in their assistive technologies are pre-trained, a typical setup for most assistive technologies. We break this paradigm by also considering blind users as active directors of their assistive technology, where via teachable interfaces blind people have an opportunity to define what is being recognized, how is it labeled, and more so, fine-tune the computer vision model with their own data. Last, when possible, we bring in broader discussions pertaining to ***privacy concerns*** accounting for both blind end-users and bystanders that may be captured by the cameras' in their assistive technologies.

1.1 'Application' Dimension

Computer vision enables cameras to support blind users in navigating their visual surroundings. Detecting and recognizing objects and people is a daily challenge for blind people [5], and thus it has been long explored in accessibility. Recent achievements in object detection and recognition with deep learning [6,7] show its feasibility in detecting many different objects, given enough training data. At first glance, it seems promising in that it could detect objects for blind users. However, detecting every single object in a scene may not be helpful to blind users who

are looking for a very specific object. In other words, the state-of-art object detection can detect as many objects as configured in the training phase, but would be unable to know which object is of interest to the user among all detected objects. The detection of the object of interest is crucial, especially in the context of machine teaching, where users can provide data to teach a machine to understand their intents (*e.g.*, specific objects of interest) [4, 8]. For example, when blind people use a teachable object recognizer [4] in real-world environments, they may often include not only a target object but other objects in the background. It would make it difficult for the machine to learn "the target object" if the machine did not know which object is of interest to the user. Observing blind people's tendency to use their hands to locate the object in the camera frame [4], we explore hand-guided object detection in the context of machine teaching in this dissertation (Chapters 3, 4, and 6.

The emergence of deep learning in computer vision also leads to growing interests in the application of face detection and recognition [9]. Face detection typically produces feature vectors that distinguish one person from others. The recent data-fueled method specifically facilitates this feature map extraction, which enables more fine-grained, robust face recognition than before. Some prior work has taken this advantage by employing such facial recognition methods in assistive technologies directly [10, 11]. In addition, facial embedding data are shown to be feasible to extract visual attributes, such as age/gender [12] and eye gaze [13]. Prior work shows that people with visual impairments [1] report that access to these visual cues can make them more active in social interactions [14]. In Chapters 6 and 8, this dissertation explores methods that can help blind people access these visual attributes via head-worn cameras, with a focus on societal issues

[1]There is debate on whether to use identity-first language or person-first language who describing a person who has a disability. We use both throughout this dissertation to respect different opinions.

that such wearable cameras may raise [15, 16].

1.2 'Camera Form Factor' Dimension

The widespread deployment of smartphones has increased access to the camera and hence computer vision technologies employed for supporting the use of such mobile cameras. This increased availability of cameras enables assistive technologies to exploit recent achievements in computer vision. However, it is challenging for users with visual impairments to aim the camera toward the area of their interest, as they have limited access to the spatial layout of the camera frame [17]. Different camera aiming behaviors change visual characteristics of images or videos from the user's camera and thus affect the performance of computer vision models. Thus, to assist camera aiming of users with visual impairments, prior work has proposed various approaches that convert the visual layout information into non-visual feedback [18–20]. In earlier chapters of this dissertation (Chapters 3 and 4), we investigate blind users' interactions with hand-held cameras (*e.g.*, smartphone cameras) with and without non-visual feedback for camera aiming.

Camera aiming behaviors can vary by its form factor. For example, while blind users need to move their hand-held camera for camera aiming, they may show different behaviors with wearable cameras. The camera form factor can also change the visual characteristics of data. Recently, many assistive technologies have started exploring head-worn cameras [21–23], which allow users to have hands-free interactions with the cameras. These efforts have resulted in commercial smartglasses products—for example, Aira [24] and OrCam [10]. Despite the earlier work and products, little knowledge is available to help us understand blind users' interactions with such wearable cameras. For example, do blind users successfully capture the area of their

interest with wearable cameras? If not, are there any patterns in their camera aiming or captured data that lead to design implications for better camera aiming? We delve into the potential and challenges of head-worn cameras for blind users in the context of pedestrian and object detection.

1.3 'User Role' Dimension

When people are using a system, they can either actively be engaged in the system or passively consume outputs from the system. Some systems require users to have active involvement. For example, the paradigm of machine teaching engages users in more active interactions by having them teach their machines to learn about their intentions, which we call *active director of technology* as initially defined in prior work on teachable interfaces [25]. Employing the concept of teachable interfaces in assistive technologies [3], we consider blind users' active engagement in teachable object recognition. In this context, the users are supposed to provide data to train their systems. We believe that this active user engagement would allow us to increase the awareness and understanding of accessibility problems in real-world situations and understand the potential and challenges in datasets for accessibility. On the other hand, in some assistive systems, users simply consume the system's feedback, which we call *passive consumer* [25]. This dissertation also explores the users' passive interactions in the context of pedestrian detection. For this study, an intelligent head-worn camera is provided to help them access pedestrians' visual attributes, such as age and gender, and proxemics signals, such as distance, position, and head pose.

1.4 'Privacy Concern' Dimension

The *always-on* nature of cameras in assistive technologies may lead to societal issues, such as ethical and privacy concerns, as these cameras can *capture* visual information pertaining to others without intention or permission. Prior work studies the social acceptance of such assistive cameras but focuses on either blind users or potential bystanders. Exploring assistive technologies that require users to make certain gestures for interactions, researchers often focus on the perspective of users to study their perception and acceptance of such technologies [26,27]. When assistive technologies use camera devices, the perspective of bystanders often takes priority over the user's as the camera can capture bystanders, which may invade their privacy. Profita *et al.* specifically tackled this issue to understand the social acceptance of assistive wearable cameras from the perspective of bystanders. Such camera-based assistive systems may include unwanted or private information of blind users, but there is no way for them to check if their data accidentally contains such information. It would also affect the social acceptance of assistive technologies as blind people do not want to reveal any information without their permission [28]. In this dissertation, we consider both the perspectives of users and bystanders in order to conjunctionally study privacy concerns and social acceptance of camera-based assistive technologies.

1.5 Overview of This Dissertation

Chapter 2 provides the necessary background, which motivates and informs this dissertation. The following chapters, comprising Chapter 3 through Chapter 5, focus on object detection and recognition, mostly considering *hand-held cameras*, the most common form factor for assis-

tive technologies that employ computer vision. More specifically, Chapter 3 presents our initial efforts in dataset collection and discusses the feasibility of a computer vision model for localizing the object of interest in photos taken by blind people based on the user's hand proximity to the object. Chapter 4 describes a sonification method to guide blind users in accessing the object-center position within the camera frame when training their teachable object recognizers. This sonification method builds on the machine learning model first presented in Chapter 3 for object localization. While Chapters 3 and 4 consider only one form factor: hand-held cameras, Chapter 5 broadens our approach around hand-guided object localization to both *hand-held and head-worn cameras*. More so, it introduces a new machine learning model that incorporates the hand information in detecting the object of interest more accurately.

The next chapters starting from Chapter 6 to Chapter 8 focus on *head-worn cameras*. In Chapter 6, we revisit the sonification method, which we initially designed for hand-held cameras and further improved to be more accurate. We deploy this approach in head-worn cameras in the homes' of blind participants; the application context here remains the same, teachable object recognition. It is only in Chapter 7 and Chapter 8 where we switch to a different application context, *pedestrian detection* while fixing the form factor to be head-worn camera. Specifically, in Chapter 7, we seek to understand how blind users interact with smartglasses in the context of pedestrian detection. Chapter 8 investigates societal issues and concerns, which assistive smartglasses may bring, from two different perspectives—those of blind users who wear and use the smartglasses and those of sighted pedestrians who may be captured and detected by the assistive application. Last, Chapter 9 concludes this dissertation by presenting its contributions and discussing research directions for future work.

Through out the dissertation, the use of the plural "we" is used to indicate the collaborative

nature of this research involving both my advisor, other faculty that provided guidance, fellow researchers and students, as well as the community participating in the studies. In the subsequent chapters, I indicate via footnotes those contributing and the specific manuscripts submitted as joint work, which are reflected in those sections.

Chapter 2: Related Work

This chapter discusses prior work to better contextualize and provide background and motivation for the work in this dissertation, which draws upon on the larger interdisciplinary literature on computer vision (*e.g.*, in the context of assistive technologies) human-computer interaction (*e.g.*, in the context of blind people's interactions with cameras). Specifically, Section 2.1 summarizes related work using computer vision for assistive technologies, focusing on the detection and recognition of objects and people as well as teachable interfaces. It also describes the efforts in filling the gap in employing computer vision in accessibility motivating the work in Chapter 3, which focuses on one of the pressing challenges, the lack of data sourced from the blind community; these are datasets that can be used to train and benchmark machine learning models in AI-infused assistive technologies. Section 2.2 covers prior work in understanding blind people's camera interactions for both hand-held and head-worn cameras. Observations from these prior studies informed our sonification method for object localization in Chapter 4 and motivated a novel machine learning model in Chapter 5. We delve deeper into the potential and challenges of wearable cameras for supporting blind users in accessing visual surroundings, literature that informs the use of this form factor for recognizing objects (Chapter 6) and people (Chapter 7). More so, we discuss how this wearable camera form factor has attracted researchers' attention in its social acceptance due to the always-on nature of its camera. We build upon this work in

Chapter 8 further investigating the societal issues and concerns that it may raise.

2.1 Computer Vision for Blind People

Assistive technologies, especially for people with visual impairments, have employed computer vision algorithms to help the users access the visual information that surrounds them. In particular, egocentric vision models are employed in assistive technologies as people with visual impairments often use their smartphones or wearable cameras to take photos of their target. Accessibility researchers have proposed various approaches to help people with visual impairments capture their surroundings and access that information. Such technologies, called *blind photography*, have been developed to simply capture scenes (*e.g.*, [29]), and recognize objects (*e.g.*, [17, 20, 30]) or people (*e.g.*, [11, 19]). In particular, prior work on blind photography reported that people with visual impairments are vulnerable to false-positive object localization [31] as there are no means for these users to check the localization output.

2.1.1 Object Detection and Recognition

Researchers have explored assistive technologies that use camera input to help visually impaired people identify surrounding objects. The following surveys prior work to identify the characteristics of these solutions and to understand: the diversity of user input, the mechanisms underlying prediction, and sources of training examples for deployed computer vision models. Table 2.1 presents representative examples from 2010-2018 focusing mainly on real-world applications, though similar patterns may be found when examining larger surveys [48–50]. While a few examples are similar to barcode scanners, requiring adhesive tags to be attached to objects,

Table 2.1: Input, model, and prediction characteristics in prior object recognition for the blind.

System	Input		Trained Model			Prediction	
	img/video	tag	sighted	blind	personal	human	machine
VizWiz [32]	•					•	
Digit-Eyes [33]		•		•	•		•
EyeNote [34]	•		•				•
LookTel Rec. [35]	•				•		•
TapTapSee [36]	•		•			•	•
Zhong *et al.* [30]	•		•				•
Talking Goggles [37]	•		•				•
CamFind [38]	•		•				•
EyeSpy [39]	•		•				•
BeMyEyes [40]	•					•	
BeSpecular [41]	•					•	
Aipoly Vision [42]	•		•				•
Kacorri *et al.* [4]	•		•	•	•		•
Sosa-Garcia *et al.* [43]	•			•	•		•
Aira [24]	•					•	
Seeing AI [44]	•		•				•
VocalEyes AI [45]	•		•				•
WayAround [46]		•		•	•		•
Envision AI [47]	•		•	•	•		•

(*e.g.*, [33, 46]), the general trend is to use camera stream as an input for prediction with a single photo [34, 36–39, 41, 44, 45, 47] or real-time video [24, 35, 36, 38, 40, 42, 44].

Prediction using computer vision models is common practice (Table 2.1). However, current limitations in object recognition make it impossible to build a "super" object classifier to recognize all possible object instances of interest for all blind people. Thus, such solutions are typically restricted to a few object instances (*e.g.*, US currency [34]), adhesive tags (*e.g.*, [33,46]), objects whose images are available on the web or a database (*e.g.*, [30, 37]), and generic object classes (*e.g.*, [38, 42, 44]). To overcome some of these limitations, applications are either backed up (*e.g.*, [36]) or fully supported (*e.g.*, [24, 32, 40, 41]) by human respondents.

Most recently, researchers are exploring teachable object recognition which allows users to personalize a recognizer with their objects of interest. Early approaches required training examples provided by sighted users [51]. Subsequently, models trained by users with visual impairments have shown promise given well-framed training examples [4, 43, 47]. As there are few datasets with photos of objects taken by blind users, most models are trained on photos taken by sighted people. Such photos may exhibit different background clutter, scale, viewpoints, occlusion, and image quality than those taken by blind users at prediction time.

Informed by this design space, we explore the presence of users' hands in photos during **prediction** for a crowdsourced solution, such as VizWiz [32], and a teachable solution, such as Sosa-García and Odone [43]; they released their datasets (Section 3.2.2). Moreover, we explore the effectiveness of our hand-guided object localization in the context of **training**, by replicating prior work from Kacorri *et al.* [4] on a new small but rich benchmark dataset (Section 3.2.2.6).

2.1.2 Teachable Object Recognition

Given the need for fine-grained labels, object recognizers typically rely on the presence of barcodes [10, 44, 52, 53], readable text [10, 44, 54], remote sighted help [24, 32, 40, 41], or large training datasets [36–38, 42]. Each method has its pros and cons. Barcode readers assume the presence of a barcode that is included in the database. Not all objects include text and product labels are often not readable. Remote sighted help can have high recognition rates given a well-framed photo but can be slow as it assumes crowd availability 24/7, often comes with a per demand cost, requires a good data plan, and more importantly raises privacy concerns.

Applications that use computer vision can mitigate some of these concerns. However, fine-grained object classification tasks often face challenges of their own such as lack of training data, a large number of classes, and high intra-class versus low inter-class variance [55]. A workaround is to significantly constrain the recognition task to a specific object category such as money readers [44, 56] or a specific user such as teachable object recognizers [4, 43, 47, 51].

By constraining the recognition task to a specific user, teachable object recognizers both limit the number of classes and reduce variability between images used to train the models and those taken for recognizing real-world objects, which are subject to similar conditions and idiosyncratic characteristics of the user [4, 57]. However, these approaches are susceptible to camera manipulation challenges faced by people with visual impairments [2, 4]. This is why some of the first attempts in teachable object recognition by Sudol *et al.* [51] required the presence of sighted help in taking high-quality training photos, which limits users' independence. Subsequent work in this direction by Kacorri *et al.* [4] as well as Sosa-García and Odone [43] focused on empowering people with visual impairments in training their own object recognizers without sighted

help, a focus that this work shares. While still in an experimental phase, we have seen that some of these attempts are reaching real-world applications for training on personal objects [47] and faces of familiar people [10, 44, 47]. Similar to the use of adhesive Braille labels or other ad-hoc approaches, the training phase in these applications assumes that people with visual impairments know the label of the object of interest at some point (*e.g.* when first obtaining it).

One of the main challenges in helping people with visual impairments train their object recognizers is automating feedback for high-quality training examples. Perhaps this explains why teachable recognizers are initially deployed for personalizing barcode readers [10, 33, 46] or faces in scene description [10, 44, 47]. In both cases, we know what to look for (a barcode or a face), and can utilize pre-existing computer vision approaches and rich datasets for those recognition tasks. This is not the case for object recognizers addressed in this work. The shape of an object of interest for a specific user (*e.g.*, a keychain, an artisanal product, or art project) is not known a priori, nor is the perspective with more distinguishable characteristics for that object. More so, the presence of multiple objects in the frame makes it difficult to know which is the intended object.

2.1.3 Person Detection and Recognition

Blind people generally have very limited sources to access information about other people in their daily lives. This hurdle consequently leads them to difficulty in initiating social interactions where visual cues, such as gaze, are essential for social initiation. Many assistive systems have been developed to help blind people obtain visual cues from other people or recognize their family members and friends [10, 11, 17, 58–62]. Early work focused more on helping blind and

low-vision people take photos of their target [17]. From a photo, the proposed system detects human faces and informs of the faces' locations via nonvisual feedback, but this system is not incorporated with face recognition. As computer vision algorithms evolve, face recognition has been employed in assistive systems by using either WiFi-enabled glasses with a camera [58] or a smartphone [59]. The studies claim that their systems help blind people recognize people without evaluation with potential users, raising questions about feasibility and usability of such systems. To address these questions regarding assistive wearable camera systems designed for blind people, we build a prototype system on smartglasses and evaluated it with blind people.

There is prior work that has explored real use cases of such assistive systems. Zhao *et al.* evaluated an accessibility bot of Facebook messenger on a smartphone with six participants with visual impairments [11]. The accessibility bot not only recognizes names but also provides other information, including facial expressions and attributes, of a person if the person is a friend of the user. However, the results of participants' actual use cases of this system were very diverse (even among the six participants), making it difficult to extract common patterns in their interactions with the system. In our work, we focus on a single use case, *pedestrian detection*, of an assistive wearable camera to limit variations and to understand people's interactions with the wearable technology in depth. Moreover, another work using multiple cameras (two stationary cameras and two wearable cameras) explored various characteristics — such as camera placement, field-of-view, and image distortion — in capturing and detecting people [61]. The authors highlight that using a wearable camera would be necessary, as it provides egocentric spatial reference to users. Bystanders (non-users), however, might have different perceptions of such wearable cameras depending on the form or visibility of the cameras. To further investigate this aspect, we use two wearable cameras (a GoPro camera and a pair of smartglasses) that vary in camera

visibility in Chapter 8.

2.2 Blind People's Camera Interactions

The availability of a camera has increased since the emergence of smartphones, and the recent breakthrough in computer vision enables camera systems to interpret imagery data more accurately. It results in more opportunities for blind people to use a phone camera as a tool for accessing visual surroundings (*e.g.*, object recognition [2, 4, 29, 32], people recognition [11, 17, 19]). In particular, researchers have tried to understand how blind people would be interacting with their phones when using its camera to recognize objects. Prior work shows that blind people tend to use their hand to localize an object while using the other hand to hold a camera phone and take a photo [2, 4]. However, camera manipulation on this form factor can be challenging to blind people, which leads researchers to explore non-visual feedback mechanisms that can help them aim a camera toward the region of interest [11, 17, 29, 31, 32].

Due to the feasibility of providing hands-free interactions, wearable cameras have gained attention from not only researchers but also product designers in assistive technologies. For example, OrCam [10] designs head-worn cameras that help people with visual impairments or blindness read a text and recognize their friends and family. Although wearable cameras do not require wearers to aim a camera with hands, camera aiming with such cameras could still be challenging to blind people due to its dependency on the wearer's body or head pose. Prior work demonstrates that the wearable cameras' position can lead to different characteristics of captures from sighted people [63] and blindfolded sighted people [61], which might be observed from blind people using wearable cameras. In particular, as blind pedestrians tend to veer [64],

this veering tendency would eventually affect the field of their camera view if they are wearing a camera. Thus, to help blind pedestrians navigate toward their destination, Fiannaca *et al.* suggests using audio feedback for a head-worn camera [21]. Some prior work tries to use head-worn cameras to aid people with low vision [65–67]. However, little work focuses on understanding how blind people would be using wearable cameras in dynamic situations such as walking. In our work, we explore the feasibility and limitations of a head-worn camera that assists blind people to detect and recognize pedestrians and objects.

2.2.1 Smartphone Camera Manipulation

We consider how challenges in camera manipulation faced by blind people have been studied in prior work. While not restricted to object recognition, prior studies analyzing photos taken by blind people have reported on difficulties related to blurriness, lighting, framing, composition, and overall photo quality [68,69]. To overcome some of these challenges, researchers have leveraged models of visual attention [70] to either provide feedback to the user for better framing [29] or to select a high-quality photo from a video stream [30]. Others have explored solutions that can detect blurriness and inform users about camera tilt [71]. While rarely discussed in the context of object recognition, such approaches assume that the users can spatially localize an object and aim the camera appropriately. Thus, they have limited utility in the case of partially included or out-of-frame objects of interest, cluttered backgrounds, and the presence of multiple objects.

Some reported strategies that blind users employ for taking good photos [17, 69, 71, 72] include: (i) positioning the camera to the center of an object and slowly pulling away while trying to keep it in frame, (ii) making an educated guess on where to point the camera, and (iii) taking

several shots, hoping some will include the object. Jayant *et al.* [17] built upon the first strategy to guide users by asking them to take a first photo of the object up-close and then to follow verbal feedback to move the camera away. Although this method is an interesting alternative approach that can be used in our work, it is verbose and has not been evaluated in a real-world setup and the presence of multiple objects.

Even though the presence of users' hands in the camera frame (mentioned in Kacorri *et al.* [4]) is not explicitly reported above, we associate it with the second strategy: users leverage proprioception [73] to make an educated guess. Prior work on blind users' fingertip detection to estimate the area of interest in a camera frame for inaccessible physical interfaces [74] or color detection [75] provides additional evidence for the potential of our approach.

2.2.2 Assistive Wearable Cameras

Since wearable computing devices became capable of obtaining visual data via built-in cameras, not only researchers [16, 21–23, 58, 60, 76–80] but also companies [10, 24, 81] have discovered the potential of wearable camera technology in augmenting the capabilities of people, especially those with disabilities. The findings can be categorized into two groups: one that is focused on users of the technology and another that is interested in the opinions of bystanders observing the technology. Prior work has investigated uses of wearable camera devices in indoor navigation [21, 79], vision aid [22, 23], and face recognition [58, 60, 61], especially for people with visual impairments. Factors for adoption of such technology are evaluated by potential users [22, 80]. While this work focuses mostly on the user's perspective, there is prior work that uses a bystander's viewpoint to evaluate wearable camera technology [16, 82]. However, little

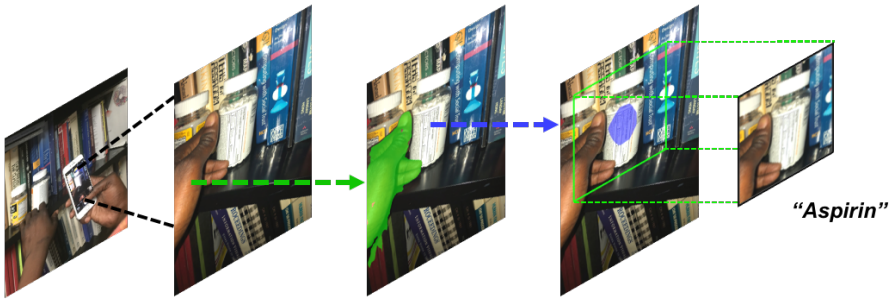
work has been done to combine both of the user's and bystander's perspectives in terms of social acceptance evaluation. We try to note social tensions between users and bystanders by looking at both groups' opinions on wearable cameras for assistive purposes in Chapter 8.

Chapter 3: Hands Holding Clues in Teachable Object Recognition on Hand-Held and Head-Worn Cameras[1]

Object recognition apps making use of built-in cameras on mobile or wearable devices have gained in popularity among blind users as they help with access to the visual world. Typically, they provide solutions employing sighted help through crowdsourcing (*e.g.*, Aira [24] and BeMyEyes [40]) or machine learning with pre-trained image recognition models (*e.g.*, SeeingAI [44] and Aipoly Vision [42]). They have also employed hybrid solutions, falling back to sighted help when computer vision models fail (*e.g.*, TapTapSee [36]). Camera manipulation, however, remains a challenge for blind users [4, 17, 29, 30]. In all cases, the effectiveness of these solutions is confounded by blurred images, viewpoints with low discriminative characteristics, cluttered backgrounds, low saliency, and more importantly *partially included* or *out-of-frame* objects of interest.

Human-powered applications, especially those employing video services, can provide near-real-time guidance for better camera aiming. However, they often come with other limitations such as the need for crowd availability and Internet connection, per demand cost, and privacy concerns [83–85]. Such limitations can be resolved with pre-trained computer vision models us-

[1]This chapter presents joint work with Professor Hernisa Kacorri published in ACM CHI 2019 [2]. Multiple graduate students at IAM Lab contributed to the TEgO dataset introduced in this chapter: Ebrima Jarjue, Dan Yang, June Xu, and Simone Pimento.



(1) User Input **(2) Hand Recognition** **(3) Object Localization** **(4) Classification**

Figure 3.1: An illustration of our hand-guided object recognition approach on an example from our egocentric dataset. Given a photo of an object in proximity to a hand, it first identifies the hand and then estimates the object center, which is then cropped and passed to the recognition model.

ing the device's onboard processing, though many of the camera manipulation challenges remain difficult to resolve. For example, a computer vision app, which is typically trained on well-framed photos of sighted people, may not be able to recognize the object of interest when given a partially included object against a cluttered background. In the case of object recognizers with teachable machines [3, 4, 43], where machine learning models are trained and personalized by blind users, having objects partially included or even out of the camera frame is a primary factor limiting performance [4]. To overcome these challenges, we are interested in examining hands as a natural interface for including and indicating the object of interest in the camera frame (as shown in Figure 3.1).

3.1 Motivation

When exploring the feasibility of teachable object recognizers for the blind, Kacorri *et al.*, 2017 [4] observed that users' hands tend to be present in many of the images, either holding an object or serving as a reference point to place the object in the photo. We speculate that they were

leveraging proprioception [86, 87], the perception of body and limb position, to coordinate their hand movements: one hand holding and adjusting the camera to the location of the other hand that was in proximity to the object of interest. Gosselin *et al.* [73] provide evidence on the ability of blind people to use proprioception to guide hand orientation and to make rapid corrections of hand orientation during movements.

Motivated by the observation above and prior literature in proprioception, we analyze existing datasets from blind users in the context of object recognition to examine the presence of the user's hand in photos. Moreover, we investigate the effectiveness of a hand-guided object localization approach for object recognition, where an object is localized based on its proximity to the hand, cropped and forwarded to the object recognition model. By using the hand as a guide to the object of interest, we not only ensure that the object is included in the frame, but we also reduce the effect that the background might have on the recognition model. This is natural since the object is cropped and the background is removed from the image, as shown in Figure 3.1. To localize the object of interest based on its proximity to the user's hand, we use convolutional neural networks (CNNs) to first train a hand segmentation model and then fine-tune it to learn to locate the center of the object in proximity to the segmented hand. The image is then cropped to include the located object of interest. We explore the effectiveness of this approach in the context of teachable machines, where the quality of the image impacts accuracy not only during the prediction step but also in training [4]. Specifically, we show that this approach could improve the accuracy of teachable object recognizers in real-world environments where objects lay in a cluttered background. With an extensive error analysis, we provide deeper insights into the feasibility and challenges of using existing egocentric datasets from the sighted population. To further research in this direction, we make our dataset as well as the output of our models for the

error analysis publicly available in `https://iamlabumd.github.io/tego/`.

Egocentric vision analyzes images providing a first-person view from a wearable camera typically mounted on the user's body or, as in our context, a mobile camera hand-held by a blind user. As users' hands in egocentric vision can provide context about intentions, actions, and areas of interest, hand recognition is an active area of research in the computer vision community (*e.g.*, [88–90]) with many applications in virtual reality (*e.g.*, [91, 92]), augmented reality (*e.g.*, [93–95]), as well as within our community [74, 75, 96–98]. To our knowledge, this is the first study exploring hand detection in the context of object recognition for the blind.

Our hand-guided object localization leverages prior work by Ma *et al.* [99], which uses a twin-stream CNN to detect egocentric activities. Specifically, we adopt one of the network streams for our recognition pipeline (Figure 3.1). However, we opt for a different network architecture using FCN-8s [100] to improve our model's understanding of hands in the first-person view. Using photos from sighted users with a head-mounted camera, we explore its potential for photos taken by both blind and sighted users using a mobile camera over a larger number of object instances.

3.2 Methods

We introduce a hand-guided object recognizer comprising three deep learning models (Figure 3.2): hand segmentation, object localization, and object classification. Our architecture is based on prior work on egocentric activity recognition [99], though we use a high-precision fully convolutional network, FCN-8s, proposed by Long *et al.* [100].

Typically, thousands of annotated images are required to train such models. Therefore, in

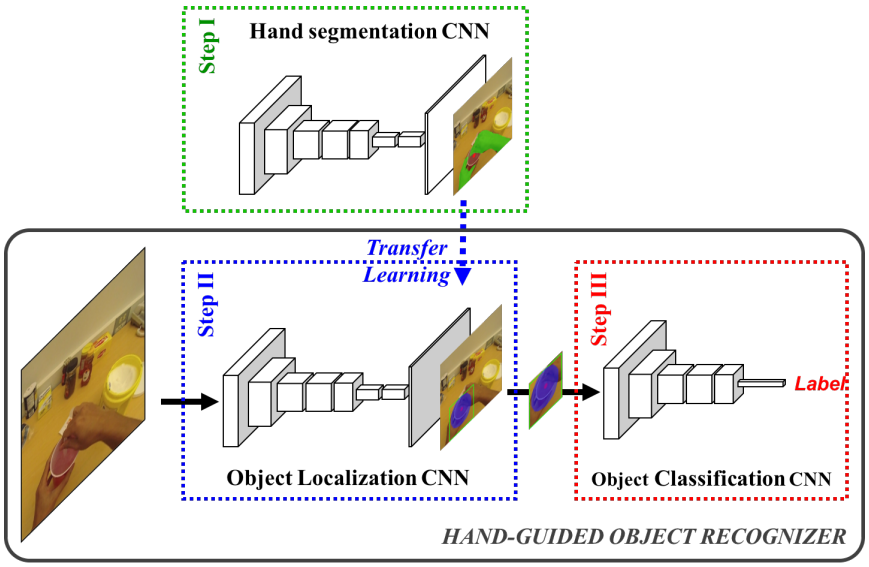


Figure 3.2: In our approach, a hand segmentation model (Step I) is fine-tuned to estimate the center of the object in proximity to the hand (Step II). A bounding box, placed in that center is used to isolate the object and crop the image, which is then passed to the object classification model (Step III).

training our system, we make use of any existing pertinent datasets (Section 3.2.2) and create a new one for richer insights and error analysis (Section 3.2.2.6).

3.2.1 Models

3.2.1.1 Hand segmentation

Our hand-guided object recognizer is based on the intuition that objects of interest will appear in the vicinity of the user's hand. Thus, to estimate the center of the object of interest, we need to locate the hand within the frame. We train a hand segmentation model to perform this task. Specifically, we use a fully convolutional network (FCN-8s), shown to perform well for fine-grained segmentation [100]. High performance is crucial to our approach as it requires highly accurate information on hand shape and poses. As shown in Figure 3.3, our model is trained with images annotated with two labels, *background* and *hand*, where we estimate the

class of each pixel; that is, whether the pixel belongs to the background or the hand. The model is optimized using the cross-entropy loss function.

3.2.1.2 Object localization

Employing transfer learning [101, 102], we build an object localization model by "fine-tuning" our hand segmentation model. Intuitively, lower layers of the segmentation model have learned to identify image pixels corresponding to a hand. Thus, these features can be re-purposed to estimate the center of the object relative to the hand pixels. Specifically, we keep weights for the first five layers in the segmentation model, shown to learn hand related features [99], and re-train the rest. Training images are annotated with a Euclidean heatmap mark indicating the object's center as shown in the left example of Figure 3.3. Heatmap annotation has been shown to be more robust than pinpointing location coordinates regarding model training [99,103]. Using cross-entropy loss, we train the localization model to predict the class of each pixel in the final layer: the background or the center location[2]. Annotating with a heatmap, the model learns which pixel belongs to the *possible center locations*. Given an image, we use this model to localize the object center and crop the image based on the estimate (blue overlay in the right example of Figure 3.3).

3.2.1.3 Object classification

Automatic object recognition spans across: (i) pre-trained models on a static number of object classes, which we call generic object recognizers (GOR); and (ii) personalized models

[2]We train both models with $10,000$ training steps, using the Adam optimizer [104], 10^{-6} learning rate, 0.9 beta1, 0.999 beta2, and 0.9 Adam epsilon.

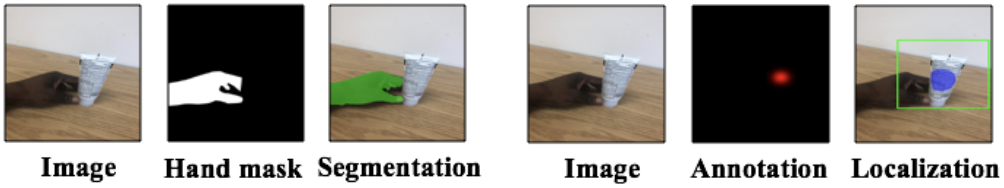


Figure 3.3: Examples of input, annotation, and output for our hand segmentation and object localization models.

with teachable machines, allowing end-users to specify objects of interest and provide a small number of training examples, which we call teachable object recognizers (TOR). We explore feasibility of hand-guided object localization for object recognition in both contexts: *generic* and *teachable*. The former uses state-of-art models trained on large data. In the teachable case, models are trained on the fly by users; thus we expect the effect of hand-guided localization to be larger as image quality affects both training and prediction.

GOR: We use Google's Inception V3 [105] pre-trained on the 2012 ILSVRC dataset [106] with 1,000 object classes.

TOR: We use transfer learning to re-train[3] GOR for 19 object classes using our benchmark dataset described in Section 3.2.2.6.

Table 3.1: Overview of the dataset characteristics used in our work. An asterisk indicates a simulated environment.

dataset	# images	# people	# objects	camera	vision	environment
GTEA [107]	663	4	16	on cap	sighted	wild*
GTEA GAZE+ [108]	$1,115$	6	-	glasses	sighted	wild*
Intel Egocentric Vision [109]	177	1	42	on shoulder	sighted	wild*
EgoHands [110]	$3,752$	4	-	glasses	sighted	wild
Glassense-Vision [43]	850	3	71	phone	low vision	vanilla
VizWiz [1]	$31,173$	-	-	phone	blind and low vision	wild
Our benchmark	$11,930$	2	19	*phone*	*sighted and blind*	*vanilla and wild**

[3]We fine-tune pre-trained Inception using $4,000$ training steps and gradient descent with learning rate 10^{-2}. We augment data by flipping left/right, randomly cropping 10% of the margin, and varying brightness by $\pm 10\%$.

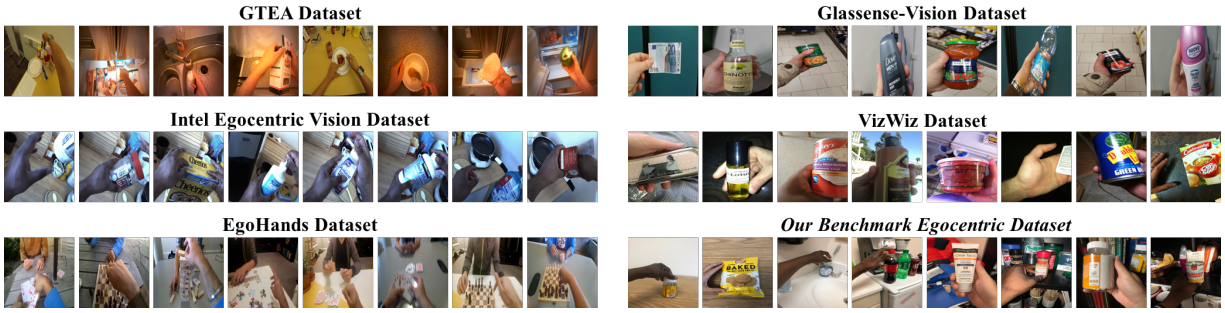


Figure 3.4: Examples from each dataset. Glassense-Vision, VizWiz, and our benchmark examples are selected to include hands.

3.2.2 Datasets

Training and evaluating each component requires an abundance of data. Ideally, the data are egocentric images from diverse groups of visually impaired people with mobile cameras to identify real-world objects using both GORs and TORs. Moreover, images require detailed annotations such as hand masks, object center, and object name (Figure 3.3). Such a dataset is not currently available.

To mitigate this, we identify related datasets, extend them with additional annotations, and create a new benchmark dataset (see Section 3.2.2.6). Table 3.1 summarizes these datasets along the number of images, distinct objects, and people, the level of people's vision, camera perspective, and environmental setup (examples shown in Figure 3.4). Environments are reported as "wild" for images captured in the real world, and "vanilla" for images in a uniform laboratory setting where objects tend to occur with a plain background.

3.2.2.1 GTEA

Georgia Tech Egocentric Activity (GTEA and GTEA Gaze+) datasets [111] are collected with 4 and 26 sighted people, respectively, through hat-mounted cameras[4]. For benchmarking activity recognition, they include annotation of actions being performed—frame-level gaze tracking data included in Gaze+ as well. They also provide hand mask annotations, since hand shape and poses are informative in object and activity recognition [107]. We manually annotated the object centers[5].

3.2.2.2 Intel Egocentric Vision Dataset

This dataset is closer to our task as it was collected for recognition of everyday objects handled by a person [109]. Similar to GTEA, two sighted participants recorded their activities using a shoulder-mounted camera. However, the default per-pixel annotations include both hand(s) and the object of interest. As our system recognizes a hand and an object separately, we manually generated hand-only masks and object-center annotations for a subset of this dataset (only 177 images from one participant)[6] and use them to train our segmentation and localization models.

3.2.2.3 EgoHands

This dataset provides first-person interactions among four sighted people; two people formed a group, and each of two wore a glass-mounted camera to record the activities [110]. We use the $3,752$ images with hand masks to train our hand segmentation model.

[4]Only data with available hand masks (6 people) were used in GTEA Gaze+.
[5]Additional annotations available at `https://iamlabumd.github.io/tego/`.
[6]Additional annotations available at `https://iamlabumd.github.io/tego/`.

3.2.2.4 Glassense-Vision

This dataset is close to both our task and our target population as it was collected with three low-vision participants for TORs [43]. Thus, we use it to evaluate our approach. A total of 71 object instances were grouped along seven categories[7] of three geometrical types (flat, boxes, and cylinders). However, their training data were collected in a uniform setting, with objects on a white surface with no hands present. In contrast, the testing data were acquired while participants held the object on their hand. Therefore, our analysis only uses the testing data that capture hand-object interactions.

3.2.2.5 VizWiz

This visual question answering dataset [1] provides a rich set of real-world images taken by visually impaired people along with questions to the crowd. While the images are not limited to object recognition, the dataset provides a unique opportunity to investigate how hand presence in real-world settings would lead to including and indicating the object of interest in the camera frame. Thus, we use this dataset to evaluate our approach.

3.2.2.6 TEgO: Our own benchmark dataset

To investigate the feasibility of hand-guided object recognition for blind people, we create a new benchmark dataset, called Teachable Egocentric Objects (TEgO)[8]. By controlling for factors such as users, environments, lighting, and object characteristics, this small but rich dataset allows us to explore and uncover potential strengths and limitations in the context of object recognition

[7]The cereals category in the Glassense-Vision dataset was unavailable at the time.
[8]TEgO is available at `https://iamlabumd.github.io/tego/`

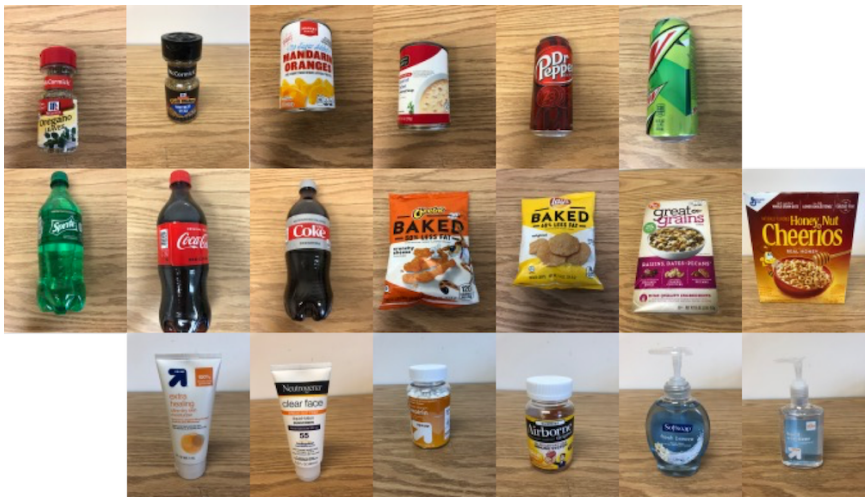


Figure 3.5: Nineteen objects used in our data collection. Objects in the same category are displayed in proximity.

for the blind, which is not feasible with the existing datasets described above.

People: Two individuals collected data over five weeks resulting in $11,930$ images. As shown in Table 3.1, the small number of people is comparable to those reported for similar egocentric datasets in computer vision, while the number of images is large. The first individual (B) was a blind undergraduate (no light perception) with little experience using a mobile phone camera. The second individual (S) was a sighted student experienced in machine learning. Data from S serve merely as *an upper baseline for discriminative power of the system given that many of the objects shared visual similarities*; that is, we treat the data from S as ideal data and use them as a point of comparison in our analysis.

Objects: Informed by prior work on TORs [4, 43], object stimuli (total of 19) were carefully engineered to cover a large group of categories with diverse geometries and functions. Objects within a category were deliberately chosen to have a similar shape and appearance (Figure 3.5). By making the recognition task more difficult, we are hoping to anticipate more challenges faced in real-world deployments.

Environment and lighting: We collected our dataset in both simulated real-world (wild*) and more traditional lab (vanilla) settings. As shown in Figure 3.1 and Figure 3.4, "wild"-photos were taken against a cluttered background including a bookshelf, a kitchen, and many other objects. Whereas "vanilla"-photos were taken against a wooden surface next to a white wall. Three different lighting conditions are considered for testing in each environment: (i) indoor lights on, (ii) surrounding light with indoor lights off, and (iii) flash on with indoor lights off. The data collection lasted for many days while the effect of natural light was mitigated.

Procedure: TORs are evaluated in two phases: training and testing. In training, the user sequentially provides a small number of photos: about 30 examples per object under each environment and lighting condition. To simulate testing, an object is randomly given to the user for a single shot. The random object assignment minimizes learning effects [4]. We iterate it until we test each object five times, resulting in a total of 19×5 test images at a time.

In the pilot stage, we observed that photos taken by B with the volume button differed from those taken with the screen button. To account for this, we also considered two photo-capturing conditions for B: (i) screen button and (ii) volume button. Since we did not observe such differences from S's photos, S used only the screen button.

More importantly, to allow for meaningful interpretation of results, data were collected under two hand-inclusion conditions: (i) use of hand-object interaction, and (ii) no hand. B took a total of $6,189$ training and $2,280$ testing images, and S took a total of $2,321$ training and $1,140$ testing images.

Annotations: Beyond object labels, all images in our dataset are manually annotated with hand masks and object center heatmaps as depicted in Figure 3.3.

3.3 Results

We use the following models for our evaluation:

Hand segmentation. The model was trained primarily on images taken by sighted people with hand mask annotations (a total of $5,707$) from GTEA, GTEA Gaze+, Intel Egocentric Vision, and EgoHands. Additionally, similar to the selective hand annotation in [111], we included randomly selected images (about 7%) from TEgO during the training phase (224 and 654 from S and B, respectively). By including some images from our dataset, we increase the potential of the model to perform well on the unseen data from S and B. It decouples the performance of the segmentation and localization models from the recognition model, and thus allows us to evaluate the utility of hands in recognition. Only the Intel Egocentric Vision dataset (177 annotated images) includes hands with a similar complexion to that of B. The majority of the data from sighted participants include hands with a complexion close to that of S ($5,530$ images). Thus, we do not anticipate that the imbalance across B and S data, will bias our segmentation model in favor of B.

Object localization. We used a total of $1,955$ images from GTEA, GTEA Gaze+, and Intel Egocentric Vision, and $4,048$ images from TEgO ($1,175$ and $2,873$ from S and B, respectively) with their heatmap annotation data to train the model; we fine-tuned the hand model to the localization model. It leads to roughly balanced data from sighted versus blind individuals ($3,130$ vs. $2,873$) and lighter versus darker complexions ($2,953$ vs. $3,050$).

Object recognition. A GOR is pre-trained on a million of images from ImageNet [106], and TORs are trained on images from TEgO; each TOR is trained with 30 examples per object (total 19 objects) from S or B in a given condition: hand presence and environment. For example,

a TOR for B is trained on 30 images per object taken in the wild with the screen button while using one hand to interact with the object.

3.3.1 Analysis on Existing Datasets from Our User Group

To provide further evidence for this work, we investigate whether hands are used as a natural interface by people with visual impairments for including and indicating the objects in the camera frame. We examine existing images from this population in the context of object recognition such as the Glassense-Vision [43] (vanilla) and VizWiz [1] (wild) datasets. In addition, we run our segmentation and localization models over these previously unseen datasets, report the proportion of images with such characteristics, and offer qualitative insights on their limitations.

3.3.1.1 Glassense-Vision

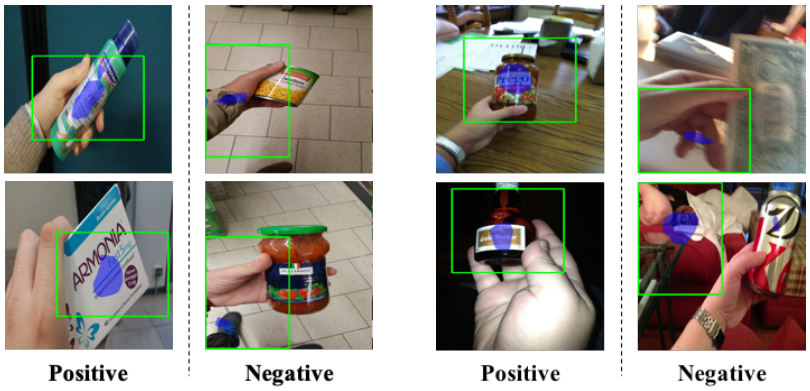
Through manual examination, we find about 44% of images taken by people with low vision in the Glassense-Vision included the individual's hand when testing the performance of their teachable object recognizers in a vanilla environment. However, our model estimated that only 16% of images include a hand, appropriately cropping the object in 89% of these. By visually inspecting failed cases, where a hand was not identified, we find images with different ratio and orientation from those typical in our training examples. It highlights the importance of diversity in hand-segmentation training data for this population. Positive and negative examples of model outputs are shown in Figure 3.6(a). While this analysis would benefit from a more quantitative approach for object detection, it requires ground-truth annotations which are currently unavailable for the dataset.

3.3.1.2 VizWiz

As discussed in Section 3.2.2, VizWiz provides a richer, more realistic object recognition scenario as blind or low vision real-world users post images with associated questions to the crowd. To gauge what portion of VizWiz data are object recognition tasks, we note that most questions begin with "what". Based on prior analysis on this dataset [1], we estimate that around 29% of questions are related to object identification: about $8,700$ of $31,173$ images. While it is impractical to manually inspect all $31,173$ for the presence of a user's hand, we run our object segmentation and localization models on the VizWiz data and detect a total of $1,548$ images that include a user's hand for the object of interest. While still an estimate, this result indicates that, in at least 18% of images regarding object identification, real-world end users use their hands as a natural interface to indicate the object of interest. Figure 3.6(b) depicts examples of our localization model's positive and negative results on VizWiz. By visually inspecting $6,000$ random samples not detected by our hand model, we identify only 36 images ($<1\%$) where users' hands were present. The majority (24) includes just the fingertips, and one image includes a previously unseen hand shape. Some images were blurry (9), and some had the camera flash on (9)—a lighting condition not present in any of the training sets. When visible, objects' shapes were: relatively flat (*e.g.*, paper, currency, gift card, newspaper, DVD cover), roughly cylindrical (*e.g.*, pen, cigarette, cans, bottles, mug), and nearly rectangular (*e.g.*, smartphones, laptop, keyboard).

3.3.2 Analysis on Our New Benchmark Dataset TEgO

For our analysis, we first separate the S data from the B data in TEgO, then divide their data into three subsets based on the method applied to the training and testing images—cropped-



(a) Glassense-Vision (b) VizWiz

Figure 3.6: Positive and negative outputs of our object localization model on the Glassense-Vision and VizWiz datasets.

object (*CO*), hand-object (*HO*), and object (*O*). The HO and O methods include original images taken by S and B. *HO* contains images where S and B are holding the object, while *O* contains the rest. *CO* consists of HO images cropped by the object localization model; that is, the CO images are extracted from the HO images.

3.3.2.1 Model performance

We explore the potential of the hand-guided recognition approach in the context of teachable object recognizers by comparing recognition performance of TORs trained on the CO images to those trained on the original HO and O images, respectively. Each model is trained on around 30 images per object at a given condition (*e.g.*, environment, button), and its accuracy is calculated on five testing images per object (total of 95) in the same condition. The average accuracy across models in HO, CO, and O is reported in Figure 3.7, with error bars denoting standard error across multiple sets of testing results for S and B—12 sets for S and 24 sets for B (due to the button settings).

As expected, the models from S, serving as an upper baseline and trained on images taken

by a sighted machine learning expert, outperform those of B. They also indicate that this 19-way classification task is challenging, with CO achieving an average accuracy of 92% and an improvement of 5% and 6% on average over HO and O, respectively. Models from B follow a similar pattern, with CO achieving an average accuracy of 71% and an improvement of 6% and 9% on average over HO and O, respectively. However, overlapping error bars indicate that these differences may not be significant.

3.3.2.2 Effect of environment

Figure 3.8 shows a breakdown of the models' performance across vanilla and wild environments. As expected, overall model performance is lower in the wild, where cluttered backgrounds tend to be present. In the vanilla environment, we observe that, in general, models trained on images where S and B hold the objects (HO and CO) perform better than those where they don't (O). However, the utility of our approach seems more pertinent to the wild environment, where CO achieves, on average, 67% accuracy for B and improvement of 12% and 14% over HO and O, respectively. These results highlight the potential of our approach, given that photos of objects taken by people with visual impairments in the real world tend to include cluttered environments, as illustrated with the VizWiz dataset (Section 3.2.2).

3.3.2.3 Effect of sample size

We explore the potential of our approach for training with highly limited sample sizes of 1 and 5 with k-shot learning in Figure 3.9. Similar to Kacorri *et al.* [4], we observe that, on average, the model performance increases with the sample size. More importantly, we note that

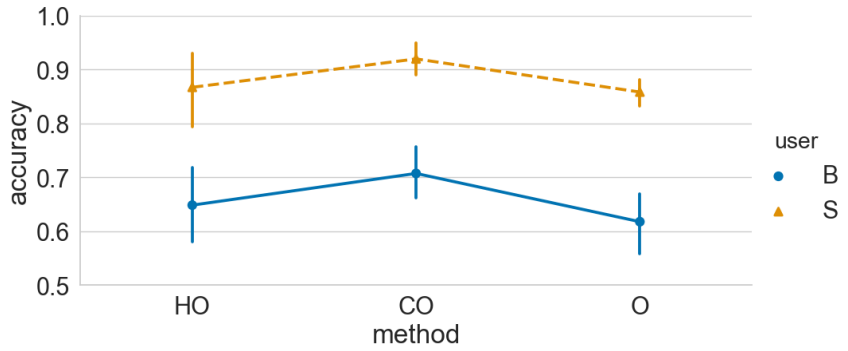


Figure 3.7: Our hand-guided object recognition method (CO) tends to improve recognition accuracy on average for S and B compared to the original HO and O methods.

CO tends to outperform HO and O consistently across sample sizes.

3.3.2.4 Teachable vs. generic

We explore whether the positive effect of the CO method over the HO and O methods, observed in teachable object recognizers (TORs), carries on to a generic object recognizer (GOR). As discussed in Section 3.2, we use as our GOR the Google's Inception V3 model [105]. Since a GOR is not trained on the labels of TEgO, its accuracy score should not be compared with those of TORs, directly. As shown in Figure 3.10, to allow for the comparison, we use V-measure [112] by comparing desirable properties of the two, such as consistency of their predictions given images of the same object, and ability to distinguish between two different objects. Similar to Kacorri *et al.* [4], we observe that a TOR achieves higher V-scores than a GOR. This is not surprising since a TOR model is fine-tuned to the users, their environments, and the number of objects. However, we did not anticipate *CO* and *HO* underperforming *O* in the GOR case. We suspect that the presence of hands in the HO and CO images, as well as close-up cropped images in the CO, may have induced some confusion in the GOR.

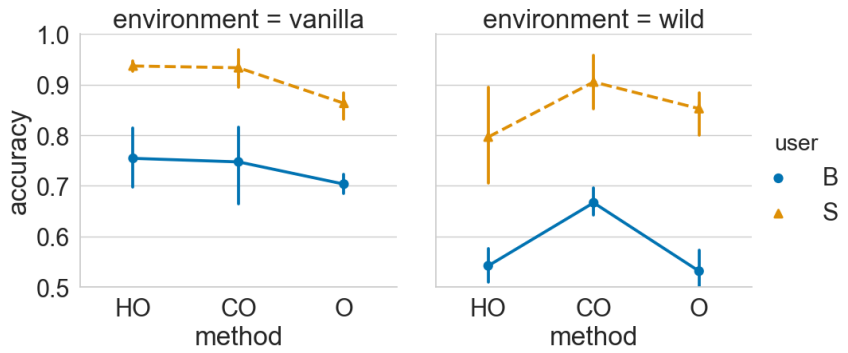


Figure 3.8: Accuracy gain of our method (CO) over HO and O is more pertinent in cluttered backgrounds (wild).

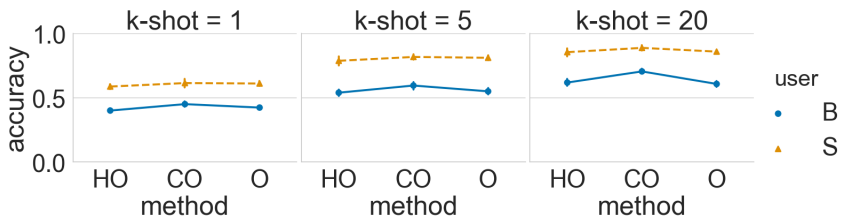


Figure 3.9: On average CO outperforms HO and O consistently across training sample sizes $k = 1, 5, 20$.

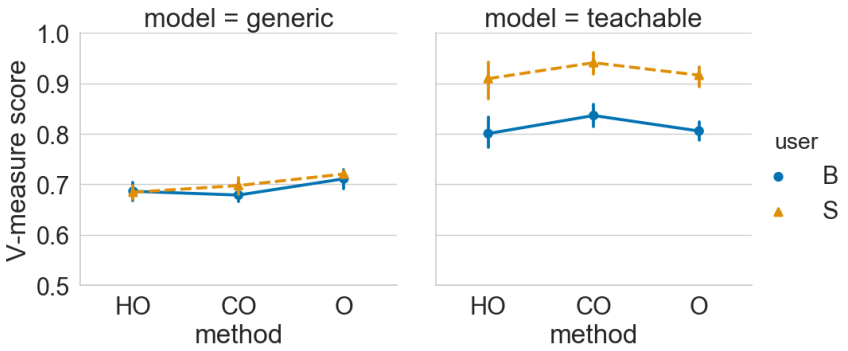


Figure 3.10: Presence of hands (HO and CO) seems to have a different effect for generic vs. teachable models.

3.3.2.5 **Error analysis**

Interested on how to improve our method, we focus on hand segmentation and localization errors (Figure 3.11) as well as object features affecting recognition (Figure 3.12).

Hands out of frame. While S and B held objects in all CO images, the B's hand is not

present in nine images; but, the objects are. These were taken in the wild environment and included relatively large objects such as soda bottles (6) and cereal boxes (3). The localization model correctly identifies object centers on seven of them and failed to localize any on the other two, which partially include the object.

Hand segmentation errors. Among 855 testing images collected in the vanilla environment, hands are partially segmented on 11 images for S and 9 for B. Even with partial segmentation, the localization model successfully infers the object centers. In 18 images from S and 29 from B, the segmentation model misclassifies small non-hand parts of the image in addition to correctly segmenting the hand. The localization model successfully infers the object centers in all but 2 images from B, where the misclassifies non-hand parts are in proximity to another object. We observe similar patterns for the simulated wild environment. Among 855 testing images in the wild, hands are partially segmented on 54 images for S and 15 images for B. However, the localization model is affected only on one image for S and two for B. In other 30 images from S and 42 from B, the segmentation model classifies small non-hand parts of the image as a hand while correctly segmenting the hand. Again, the localization model successfully infers the object centers for all but two images from B, where the misclassified non-hand parts are in proximity to another object.

Object localization errors. There are few instances in the vanilla environment where the localization model fails even though most of the hand is detected. Specifically, on two images from B, the model fails to localize the object, which is partially included. We observe 11 similar instances from B in the simulated wild environment. However, the most common localization error, observed in one image from S and 20 images from B in the wild, is misidentifying another object close to the object of interest. We suspect that some of these errors could be mitigated with

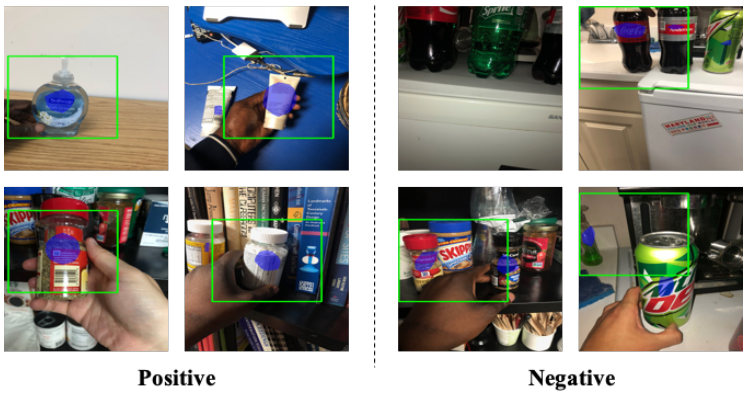


Figure 3.11: Positive and negative results on TEgO, with out-of-frame hands for some of the negative examples.

additional training examples including such ambiguities.

Discriminative features of objects. The stimuli objects in TEgO were engineered to be diverse in terms of shape and function while sharing similarities to make recognition challenging. We illustrate aggregated results on misclassified testing images as a confusion heatmap (Figure 3.12). We observe that some of the highest confusions are among objects within the same category that are cylindrical or share visual features such as soup can and mandarin can, salt and oregano, cheetos and lays. It indicates room for better performance in the recognition models (TORs) independent of the quality of cropped photos fed by our localization model.

3.4 Discussion

We discuss implications and limitations of our analysis and the proposed hand-guided object recognition method.

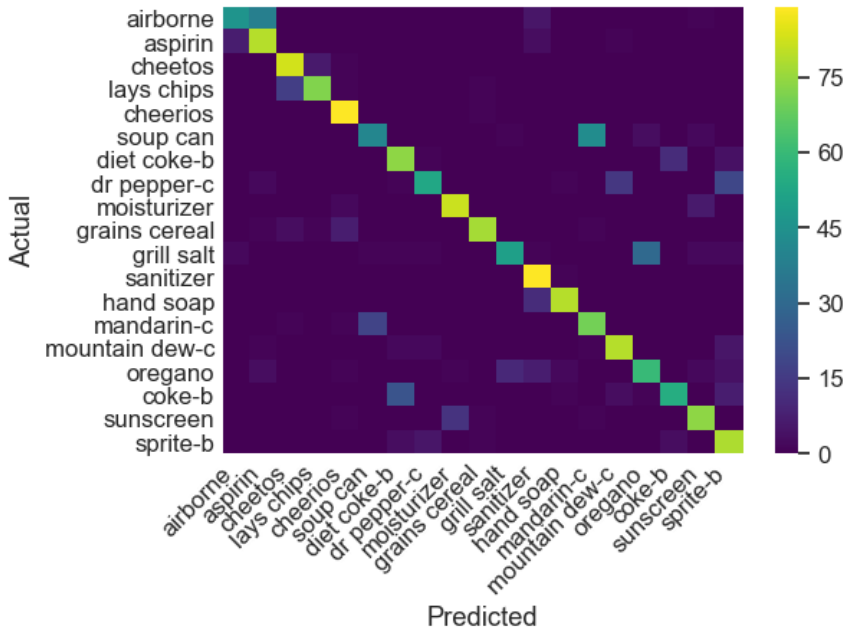


Figure 3.12: Confusion matrix for the CO models showing that misclassification occurs within objects of similar shape. Cans and bottles are indicated as "-c" and "-b", respectively.

3.4.1 Implications

Our findings and insights can contribute to the design and evaluation of future object recognition applications for people with visual impairments in the following ways.

Teachable object recognizers. We demonstrate that by jointly modeling hand segmentation and object localization, we can improve the performance of teachable object recognizers. Our analysis indicates that this gain is present across models trained by a sighted and a blind individual, vanilla and wild environments, as well as varying training sizes. The largest gain appears in the wild, highlighting the potential of this approach given that photos by people with visual impairments in the real world tend to include cluttered environments.

User interactions. Our analysis on existing datasets from people with visual impairments in the context of image recognition provides further evidence on the utility of hands as a natural

interface for including and indicating the object of interest in the camera frame. However, we observe that the presence of hands might have the opposite effect on generic object recognizers where the training images are often leveraging available data from sighted people. Given the visual feedback, sighted people may not use proprioception to aim their camera. Thus, their images may not necessarily include the hand. We propose that researchers consider these user interactions when designing and evaluating their systems.

Representative datasets. While there is a recent trend of data sharing in accessibility that we embrace (*e.g.*, [113–115]), privacy concerns often prevent researchers from sharing images taken by people with visual impairments outside the lab. This work illustrates the potentials and limitations of leveraging existing data from a non-representative population such as sighted people. Beyond a population mismatch, we faced the following challenges: bias towards lighter hand complexions; bias towards male individuals (13 male; 2 female); lack of wrinkles, tattoos, and jewerly in the hands; and limited annotations for new but related learning tasks. We contribute towards representative datasets by making our data, TEgO and other annotations, publicly available.

New directions. To provide evidence for the utility of proprioception in photo-taking, similar to the Glassense and VizWiz datasets, the blind individual in TEgO did not receive any guidance from the mobile device in taking a good photo. However, we see the potential of our hand segmentation and object localization models for providing an explicit mechanism that can guide the user. For example, we can use sonification to guide the user to take well-framed photos based on the estimated center of the object, its relative position to the frame, and confidence scores from our models.

3.4.2 Limitations

Our dataset was collected with two individuals by controlling for a number of factors which exceed characteristics of previous datasets (in terms of objects and environments). While the small number of people is comparable to those reported for similar benchmark egocentric datasets, we can see results benefiting from a larger and more diverse blind user pool (gender, age, and hand characteristics)—especially since one of the individuals was sighted and merely served as an upper baseline. Furthermore, such analysis would benefit from additional data that account for real-world cluttered backgrounds, lighting conditions, and object ambiguities.

Due to the lack of available hand data with darker complexions, we included more hand data from the blind individual in training our segmentation model. While this might have biased our hand segmentation model, we do not see that this effect carries over into the segmentation (more errors are observed on images from the blind individual). We suspect that additional data can also help improve the performance of our approach by allowing to encode diverse hand shapes in the localization better. Prior work on hand-object manipulation [116–118] argues that people tend to use same or similar hand shapes for a given object shape and function.

Given the limited data, we did not use a dynamically sized bounding box. Cropping with a fixed size box could have resulted in unwanted background artifacts or partially included objects, which may have impacted their recognition.

Due to the lack of hand annotations for the Glassense and VizWiz datasets and available tools for understanding model behaviors, part of our analysis was based on subjective visual inspection. For a more objective analysis, we are currently exploring the compatibility of available approaches, such as Grad-CAM [119] and LIME [120] for explainability.

3.5 Summary

In this chapter, we explore the utility of proprioception in the context of object recognition, which allows people with visual impairments to more easily align their camera by holding or placing their hand close to an object of interest. We provide evidence that such natural interactions occur in real-world datasets for this population and demonstrate their potential for teachable object recognizers, where out-of-frame or partially included objects against cluttered backgrounds can degrade performance. By jointly modeling hand segmentation and object localization, we achieve a sizable improvement on recognition accuracy that peeks on simulated real-world conditions with cluttered backgrounds. To train and evaluate our models, we leverage existing egocentric datasets on related tasks from sighted people and collect a new benchmark dataset (TEgO). Our extensive error analysis provides insights into the feasibility and challenges of this approach. To further research in this direction, we make our dataset as well as the output of our models for the error analysis publicly available at `https://iamlabumd.github.io/tego/`.

In future work, we plan to replicate this analysis with video streams processed at a frame level. While videos can provide continuous information on hand-object interactions, they are computationally expensive, consume more power, and require a frame-selection approach to extract higher quality images that we are still working on.

Chapter 4: Hand-Guided Sonification for Teachable Object Recognition on Hand-Held Camera[1]

Object recognition is one of the daily challenges that people with visual impairments face [5]. This is often not limited to general object categories such as shirt, soda, or medication that can be distinguished by touch or other nonvisual senses but more so to fine-grained categories such as flavor, brand, or other specific characteristics [4]. For this reason, beyond adhesive Braille labels [121] or other ad hoc organizing systems [49], people have been quick to adopt camera-based technologies that either use remote sighted help (*e.g.*, [24, 32, 40]) or leverage pre-trained image recognition models (*e.g.*, [36, 37, 42]). Proper aiming of the camera is critical for both cases as it can impact the effectiveness of these solutions.

Images that are blurred, show non-informative viewpoints, have low saliency in cluttered backgrounds, or more notably miss or partially include the object of interest (as shown in Figure 4.1) make the recognition task challenging for both humans and machines. In the case of remote sighted help, those images tend to typically slow down the response rate as crowd-workers try to provide feedback or guidance for better camera aiming [19]. However, when it comes to pre-trained models, current applications do not provide any feedback to users about the quality of their photos. Thus, low-quality images typically result in recognition errors [30], given that

[1]This chapter presents joint work with Professor Hernisa Kacorri and graduate students at IAM Lab—Jonggi Hong, Simone Pimento, and Ebrima Jarjue—published in ACM ASSETS 2019 [31].

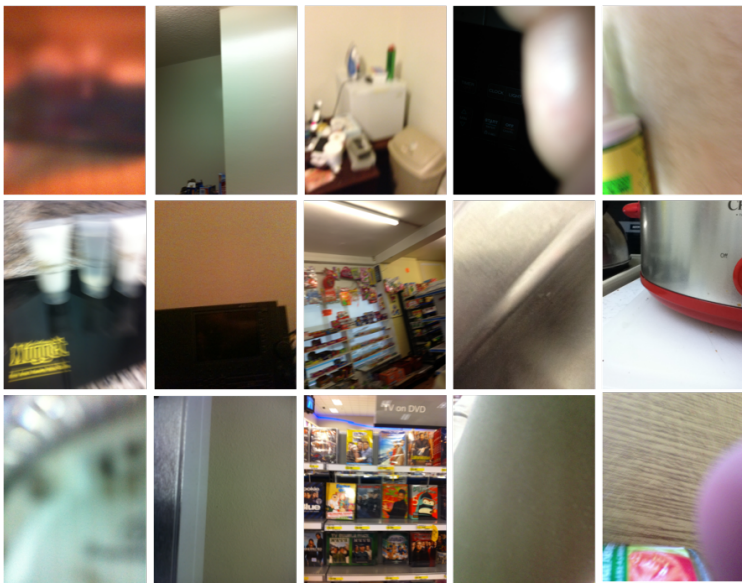


Figure 4.1: Examples of challenging photos taken by people with visual impairments sent to Vizwiz crowdworkers. About 28% of all photos are categorized as unanswerable [1].

those models are often trained with well-framed photos taken by sighted people. The problem is further exacerbated when photos from people with visual impairments are also used in training as with teachable object recognizers [4, 43, 47]; making blind photography a primary factor limiting performance.

To overcome these challenges, we are interested in exploring real-time feedback for including and indicating the object of interest in the camera frame. While camera framing guidance has been explored in accessible blind photography [17, 19, 20, 29, 122, 123], it has not been evaluated for object recognition, especially when training machine learning algorithms.

In this chapter, we introduce real-time feedback powered by a deep learning model that estimates the center of the object of interest in the camera frame. Our model is informed by prior work [2] providing evidence on the utility of hands for people with visual impairments as a natural interface for including and indicating the object of interest in the camera frame. Specifically, we use convolutional neural networks (CNNs) to estimate the center of the object of

interest in terms of its proximity to a user's hand by first training a hand segmentation model and then fine-tuning it to learn to locate the center of the object in proximity to the segmented hand. The coarse location of the object in the camera frame is then communicated in real-time through audio and haptic feedback.

We conducted a user study with nine participants with visual impairments to evaluate the effectiveness of our feedback in the context of teachable object recognition, where each participant trains a mobile phone to recognize 15 objects. We show that our feedback can help reduce training examples without the object of interest in the frame, even for blind users who have never taken a photo before. Our cluster analysis indicates that better feedback is associated with photos that include the entire object and shorter training times. While there is a negative correlation between participants' age and the performance of their models, this seems to be explained by their photography experience. Last, participants tend to trust the feedback even though they were exposed to its limitations.

This chapter's contributions are: (1) a real-time feedback approach for better camera framing triggered by natural object-hand interactions; (2) insights from a replication study on teachable object recognizers; (3) anecdotal evidence on older adults interacting with teachable interfaces; and (3) a cluster analysis method applicable to other assistive technologies.

4.1 Motivation

Our work draws from the rich literature on camera manipulation, nonvisual feedback, and object recognition for people with visual impairments. To inform the design of our feedback mechanism, we focus on sonification and haptic feedback methods previously reported for pho-

Table 4.1: Comparison of several prior applications addressing camera manipulation for people with visual impairments.

	[20]	[17, 122]	[29, 123]	[19]	[30]	[10]	[11]
Focus							
Object	•	•			•		
Barcode/Text						•	
Face		•		•		•	•
Any			•				
Feedback							
Sound	•	•	•			•	
Verbal		•	•	•			•
Haptic		•		•			
None			•		•	•	
Information							
Face Count							•
Proximity	•					•	
Directions		•	•	•			
Source							
Human	•						
Machine	•	•	•	•	•	•	•

tography and navigation. Moreover, to better contextualize the implication of our findings, we discuss current computer vision solutions for this user group with an emphasis on teachable object recognizers.

Blind photography is not a new concept. Research in this area includes efforts around the world on teaching photography to people with visual impairments [124, 125]; understanding their photography needs [11, 126, 127]; building accessible photography applications [69, 71, 128, 129]; providing remote sighted help [32, 36, 40, 41]; and, more closely related to this work, obtaining better-quality photos [10, 17, 19, 20, 29, 30, 122, 123].

We look into prior work in camera manipulation for people with visual impairments to identify characteristics of these solutions and understand the diversity of feedback modality and

information communicated to the user. Table 4.1 presents representative examples from 2010 to 2019 focusing mainly on photos for identifying objects [17, 20, 30, 122] while broadening it to other work helping with the quality of photos for reading barcode or text [10], identifying faces [10, 11, 19, 122], and capturing scenes [29, 123]. As shown in Table 4.1, when helping with the quality of the photos, the majority of solutions opt for an automatic approach, which can be instantaneous. These instantaneous quality estimates are not always communicated to the user [10, 30, 123]. Instead, photos deemed as high quality are extracted automatically from video streams.

When guidance is provided, it often includes verbal instructions such as left, up, down, right [19, 20, 122, 123] or the number of faces in the camera frame [11] though sonification and haptic feedback have been also explored. Specifically, Bigham *et al.* [20] used sonification to direct users to the target object's location. Among three different modalities (tone sound, clicking sound, and verbal instructions), they found that participants preferred the clicking sound. Similarly, OrCam [10] uses a beeping sound to indicate a good framing. This was contradicted in Vázquez *et al.* [29], where participants preferred verbal instructions to the tone sound, which only indicated object distance from the frame center. On the other hand, when Jayant *et al.* [17] only provided verbal instructions, with participants mentioning tone and haptic feedback as alternatives. While none of the prior work explored user guidance in the same context with this work, they informed the design of our feedback, exploring both sound and haptic modalities. Also, we see that communicating the position of the object relative to the center rather than just its distance can be helpful.

Nonvisual guidance is not unique to photography. For example, Gerino *et al.* [130] used a sinusoidal-wave sound and differentiated its frequency to convey spatial information. We con-

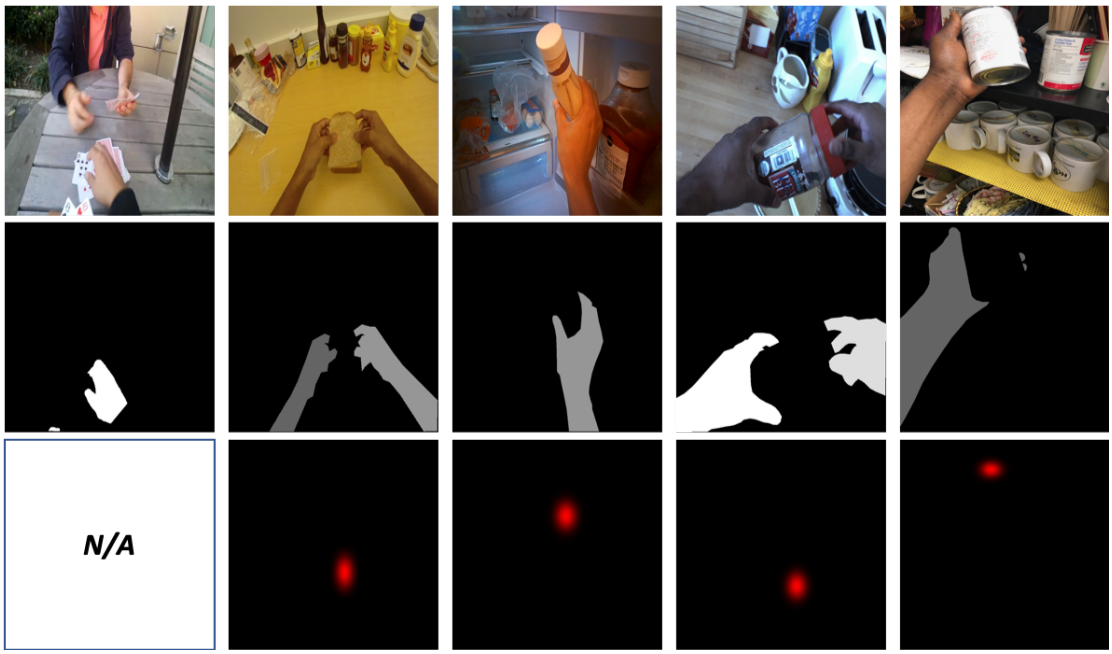


Figure 4.2: Training examples in our hand-segmentation and object-localization models using the EgoHands, GTEA, GTEA Gaze+, Intel Egocentric, and TEgO datasets (in the left-to-right order). Original images are shown on the first row, hand segmentation on the second row, and object center annotations on the last row. EgoHands is used only in hand segmentation.

sider this approach during our exploration of feedback modalities. Sound was also used by Brock *et al.* [18] to inform blind users of an object's location in 3D space; stereophonic sound was employed for x-axis, pitch for y-axis, and volume for distance from the user (z-axis). To reduce the learning cost, we consider only the x-axis stereophonic sound for our feedback, which has also been previously used in blind navigation and found to be more appropriate than verbal instructions [131].

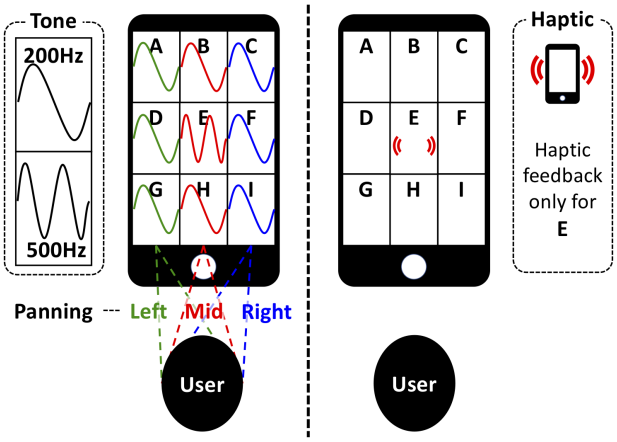


Figure 4.3: Our nonvisual feedback mechanism.

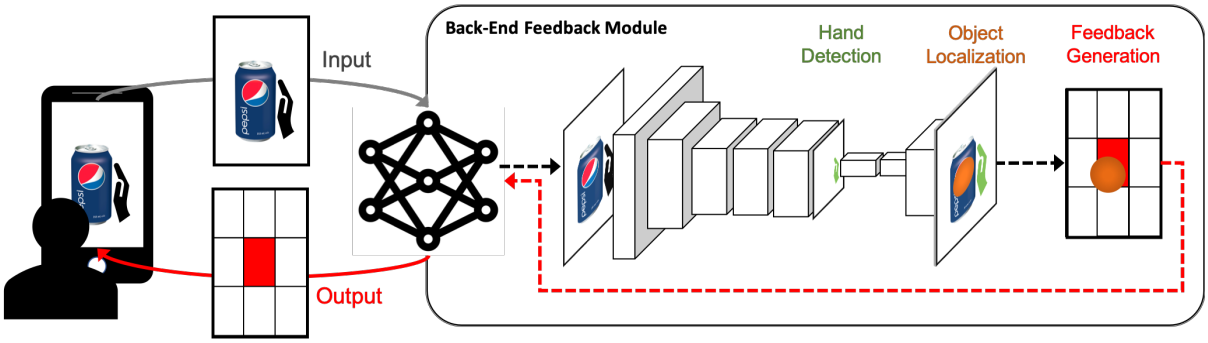


Figure 4.4: The architecture of our real-time feedback approach for better camera framing. Our testbed sends an image to the server, receives an estimated location, and triggers the corresponding feedback.

4.2 Methods

4.2.1 System Design

4.2.1.1 Hand-guided object localization for feedback

Informed by our prior work [2], we propose a real-time feedback mechanism that guides users to frame an object of interest in the camera by leveraging natural object–hand interactions. As shown in Figure 4.4, our feedback module estimates the object center location in the camera frame (object localization model) informed by the presence and shape of the user's hand in the

frame (hand segmentation model).

Hand segmentation model: We first train a hand segmentation model that identifies image pixels that correspond to the user's hands in the camera frame. We use a FCN-8s neural network architecture [100] and train it with the publicly available egocentric datasets (a total of $9,241$ training examples): EgoHands [110], GTEA [107], GTEA Gaze+ [108], Intel Egocentric [109], and TEgO [2]. All datasets focus on human interactions from first-person point-of-view: their examples in Figure 4.2 and a detailed comparison in [2]. We used the following hyperparameters in training: $10,000$ learning steps, 10^{-5} learning rate, and 16 batch size.

Object localization model: For our object localization model, we take the previously trained hand segmentation model and freeze the weights for the first five convolutional layers, shown to learn hand-related features [99]. We re-train the remaining layers with Gaussian heatmap blob annotations for object centers ($6,239$ training examples), shown in Figure 4.2. These annotations were shown to be more robust than pinpointing location coordinates [99, 103]. The hyperparameters for fine-tuning were the same as those noted in the hand segmentation model above.

4.2.1.2 Iterative design of the real-time feedback

Building on prior work in blind photography, we explore alternative feedback mechanisms that convey the estimated object center location in the camera frame. We chose our design among several options based on extensive piloting with one of the blind researchers in our team.

Explorations: Our initial attempts provided an increasingly fast and loud beep sound to communicate the proximity of the estimated object-center from the camera frame center convey-

ing limited spatial information as highlighted in Vázquez *et al.* [29], where participants preferred richer verbal instructions to the beep sound. To add spatial information, we considered stereophonic sound [18] that informs an object-center location in the x-axis with three distinguishable sounds for left, middle, and right, as shown in Figure 4.3. We piloted the stereophonic sound with a beep tone and a sinusoidal wave sound widely used in nonvisual shapes exploration for blind people [130]. We explored higher frequencies for the sinusoidal wave sound to convey a well-centered object. Also, we considered communicating the presence of the object within the camera frame through vibrations [19, 122], which were more intense when the object was well-centered. A combination of the stereophonic sound with the vibration deemed more distinguishable.

Our feedback mechanism: As shown in Figure 4.3, we coarsely divide the camera frame with a 3×3 grid alphabetically labeled from A to I. Given an estimated object-center location from our object localization model, we apply the following mapping between the region of the estimated location and the feedback modality:

- **A, D, G**: left-stereophonic sinusoidal wave (200Hz)
- **C, F, I**: right-stereophonic sinusoidal wave (200Hz)
- **B, H**: middle-stereophonic sinusoidal wave (200Hz)
- **E**: middle-stereophonic sinusoidal wave (500Hz) + vibration

The intuition behind this mapping is that users can tell when the object is within the frame, have a sense of whether it is on the left or right, and quickly tell when it is well-centered. They receive a continuous sound with changing frequency and channel, and vibrations for well-centered objects. There is no sound or vibration for undetected or out-of-frame objects.

4.2.2 User Study

To understand the potential and limitations of our real-time feedback approach for camera manipulation in the context of teachable object recognizers, we replicate the in-lab user study by Kacorri *et al.* [4], where people with visual impairments train object recognizers without any feedback on the photo quality. We further extend the study design with a simulated real-world environment and additional open-end questions.

4.2.2.1 Participants

We recruited nine participants with visual impairments from our local community (IRB #1255427-2)—each compensated for their participation at 55–80 USD (μ: 63.3, σ: 7.1). Seven participants identified as totally blind and two participants as legally blind; five of them reported having some light perception. They were all female age 29–69 (μ: 56, σ: 15) as shown in Table 4.2. None of them had used a smartphone for more than 10 years and only three (P3, P7, and P8) had previously taken photos more than once a week. Figure 4.5 shows participants' use and attitudes toward technology potentially affecting our task, based on the Rosen *et al.* [132] questionnaire.

In open-ended questions, participants reported using Braille labels on kitchen appliances, electronic devices, food items (*e.g.* cans and bottles), cups, clothes, jewelry pouch, ID/credit/gift card, and mail. When a Braille label was not available, they reported using smell, isolating the object, coming up with some ad-hoc solution such as rubber bands for differentiating texture, asking for sighted help, or using an object recognition application—though P1, P4, P7, and P9 had never used any technology for object recognition. The applications mentioned across the

Table 4.2: Participants' demographics and years of smartphone use. Light perception is indicated with an asterisk.

ID	*Gender*	*Age*	*Onset*	*Handedness*	*Smartphone*
P1	F	53	13	R	8
P2	F	69	birth*	L-reading, R-other	6
P3	F	60	birth	L-reading, R-other	8
P4	F	66	birth	L	4
P5	F	63	33	R	7
P6	F	34	28*	L	10
P7	F	29	19*	R	10
P8	F	67	birth*	L-touchscreen, R-other	7
P9	F	64	birth*	R	3

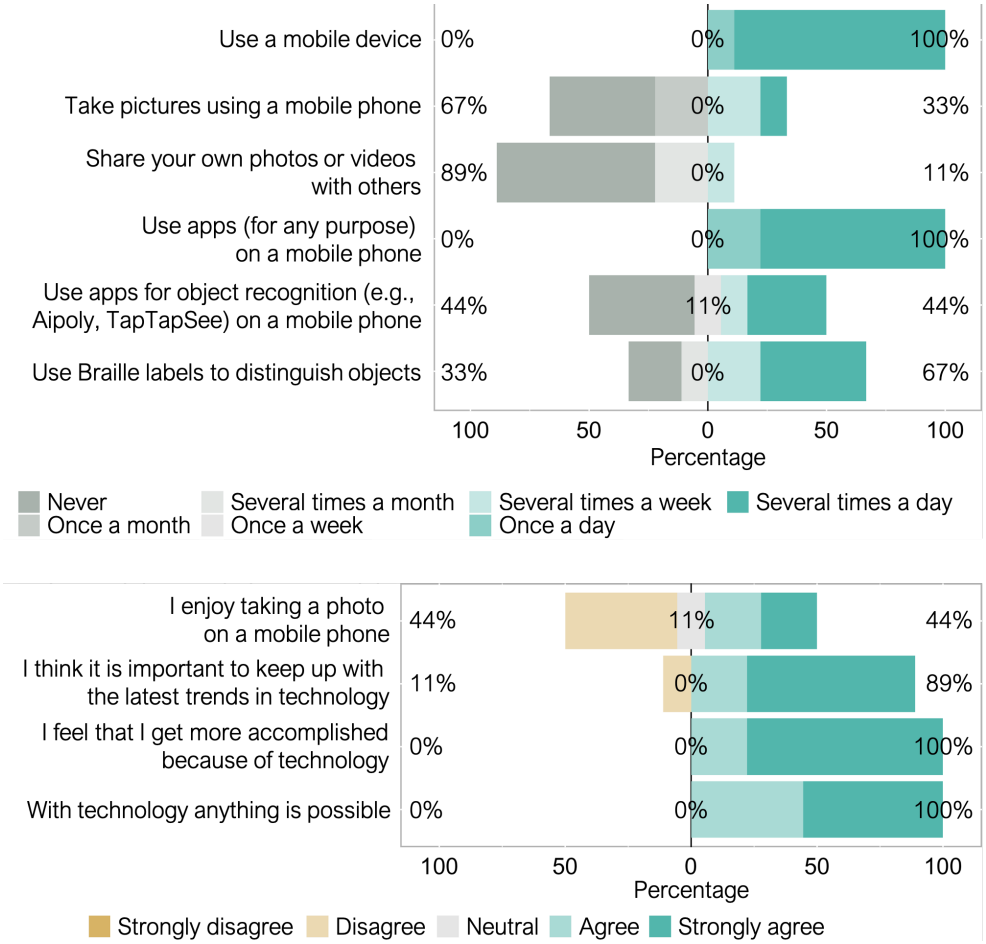


Figure 4.5: Technology experience and attitude responses. All participants have smartphones; more than half are using apps for object recognition; and all are positive about technology.

other five participants were: KNFB reader [54], BeMyEyes [40], Aira [24], and SeeingAI [44].

4.2.2.2 Testbed

We implement our feedback mechanism into a custom experimental testbed for iPhone 8, which connects over WiFi with a GPU server hosting the localization model. Every 333ms our testbed sends a resized (540x720) image to the server and receives an estimated location for the object center as shown in Figure 4.4. The estimated location is based on the image pixels classified with the highest probability as the center of the object of interest. For this study, we use $p > 0.3$ as a threshold. There is no sound or vibration when the threshold is not met (*e.g.* for an undetected or out-of-frame object). For the object center within the frame, the testbed triggers the corresponding feedback to the A–I regions. Given the high rate of 333ms, participants perceive the feedback with continuous sound changing frequencies and channels, on/off vibrations, and silent pauses, as they move the object or the camera around. Beyond the shutter sound, the testbed communicates to the participant in real-time the count of photos taken and the completion of the training process for each object.

4.2.2.3 Object stimuli

As shown in Figure 4.6(a), we use three objects for practice. The grill salt and mountain dew can are selected from the TEgO dataset [2] used to train our object localization model. Thus, the estimates for their object-center locations are anticipated to be more accurate and stable. On the other hand, the nut mix has not been seen by the object localization model, providing experience of less stable feedback. Given that our object localization model is error-prone like any

(a) a spice jar, soda can, and a snack box.

(b) 15 stimuli: baking soda, caramel coffee, cheetos, chewy bars, chicken broth, coca cola, diced tomatoes, diet coke, dill, fritos, lacroix apricot, lacroix mango, lay's, oregano, pike place roast.

Figure 4.6: Objects for (a) practice and (b) stimuli in our study.

machine learning algorithms, it is important to have participants learn about this and familiarize themselves with the fact that the feedback can be imperfect and thus it is not to be trusted at all times.

While the premise of teachable object recognizers works best for unique objects (*e.g.* key-chains or artisanal products that may or may not have readable texts or recognizable labels), researchers [4, 43] often use commercial products that allow for exploration of different shapes, sizes, materials, visual similarities, and more importantly, stimuli that allow for replicability of the study. We follow this approach and adopt the object stimuli from Kacorri *et al.* [4] (Figure 4.6(b)). Due to product changes, labels for some of the stimuli such as *k-cups*, *baking soda*, *chicken broth*, *diced tomatoes*, and *diet coke* were different, though, the shape and material that could potentially impact participant interactions remain the same for all stimuli.

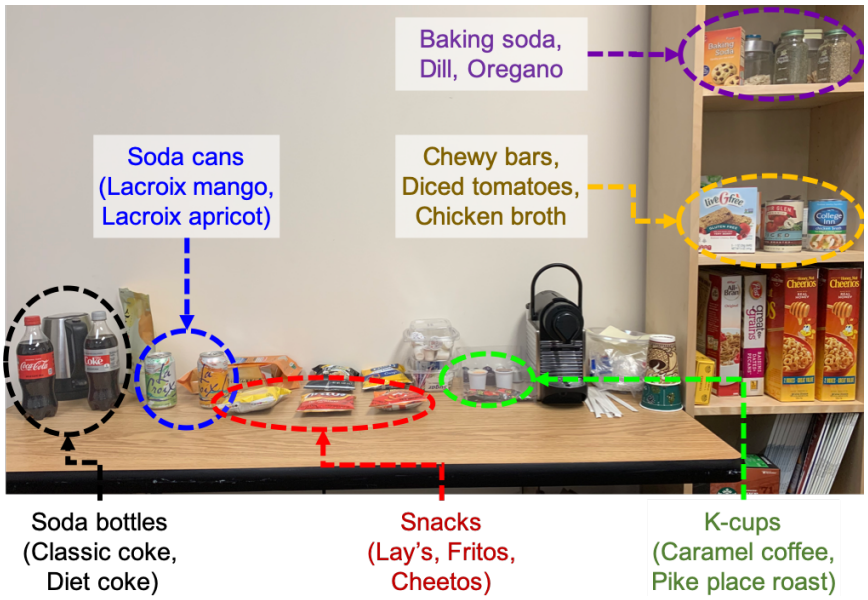


Figure 4.7: The wild test setting in the cluttered environment.

4.2.2.4 Environment

We have two environments in our study. The first is a traditional lab setup, where participants sit in front of a gray unpatterned table against a plain white wall (Figure 4.8) as in Kacorri *et al.* [4]. We call this **vanilla** as each object is placed in the plain background. To explore the potential of our feedback in more realistic scenarios, the second environment is a simulated real-world setup. We call this **wild** as photos are taken on a wooden surface against a cluttered background including a book-shelf and other objects (Figure 4.7). We control for lighting conditions across participants by having both environments indoors without natural light.

4.2.2.5 Study procedures

A study session took three hours on average. We first collected demographic information, prior experience with taking photos and object recognition apps, and attitudes towards technol-

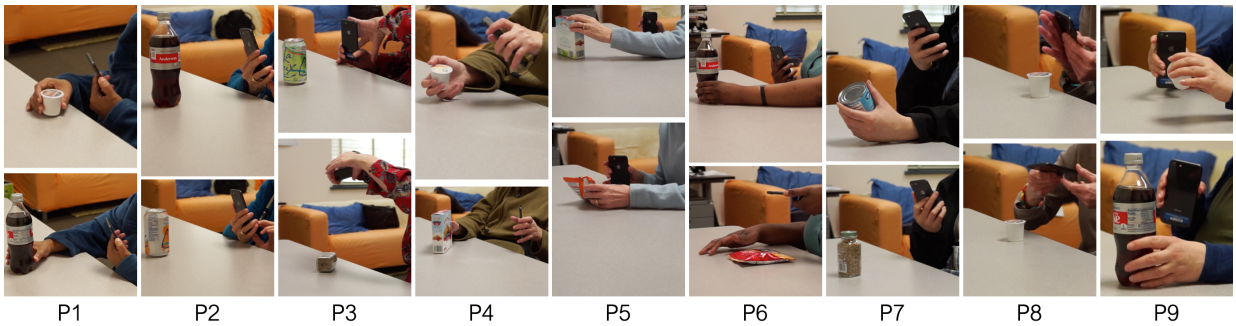


Figure 4.8: Vanilla environment and the camera manipulation across participants in train mode.

ogy. Then, participants were introduced to the study task. Using our testbed, they first learned about the feedback by taking five well-framed photos for each of the practice objects. They took photos by using either a shutter button on the screen or volume buttons on the left side of the phone.

The following instructions introduced the feedback:

> *The app tries to estimate the object of interest based on your hand; that is, given your hand position and pose in the camera, the app will try to estimate the center of the object that you are interacting with. Please note that the feedback may not be perfect; so it may play for a different object from the one you care, or may not play at all even if your object is in the camera frame. For it to work best, use your hand to interact with the objects.*

After practicing, participants took photos of fifteen object stimuli in three different modes: *vanilla train*, *vanilla test*, and *wild test*. In train mode, participants were asked to take sequentially 25 photos per object, to train their object recognizer. By hearing the count and end of training, participants provided the balanced number of examples across objects. The order of object assignment in the training mode was randomized. Participants were provided with the following instructions [4]:

- *Increase consistency.* When taking photos for training, imagine how your future self could be holding and taking a photo of that object to identify it.
- *Include object in the camera scope.* Distance the phone from the object relative to the object size (closer for a smaller object and further away for a larger object).
- *Obtain discriminative photos.* Many products tend to avoid printing their labels where the seal is. If you can tell by touch where the seal, avoid taking photos on that side.

In contrast, in test mode (both vanilla and wild), participants were asked to take a single photo of an object at a time. Objects were handed to the participants in vanilla. Whereas in wild, participants were guided every time to reach a specific object. The process was iterated five times per object in each test mode so that in the end they took five photos per object in vanilla and wild, respectively. The order of the object in each iteration was randomized to minimize learning effects.

4.3 Results

4.3.1 Object Recognition Performance

To assess the potential of our feedback approach in the context of teachable object recognizers, we used the images taken in train mode from each participant and built for each a personal object recognizer. Specifically, ten recognition models were trained per participant by randomly selecting 20 out of the 25 training images per object. To allow for comparability of our results with those of Kacorri *et al.* [4], we used the same recognition model based on Google's Inception V3 [105] pre-trained on ImageNet [106] and fine-tuned it to the randomly-selected train data of

the participant. Our hyper-parameters in training the models were: $1,000$ training steps, 100 batch size, 0.01 learning rate, and no data augmentation. These parameters are the same as those used in Kacorri *et al.* [4].

Similar to Kacorri *et al.* [4], we recruited two sighted people (S1 and S2). Their data are used to gauge the performance of the CNN architecture on the stimuli and merely serve as an upper baseline. S1 is a 26-year old male sighted computer scientist with machine learning experience. S2 is a 31-year old male sighted economist with a basic understanding of it.

Promising performance for those with no camera experience. Figure 4.9 shows the average accuracy of the participants' object recognition models on their (*vanilla test* and *wild test*) data, respectively. Participants (P1, P2, P4, P5, P6, P9), who had none or little experience of photography before the study, achieved at least around 50% accuracy on average in the vanilla test. By comparison, a random 15-way classification would yield about 7% accuracy. We observe that the models from our participants performed comparably to those reported in Kacorri *et al.* [4], even though our study includes a participant pool that is 16 years older on average with limited experience both in blind photography as well as mobile apps that use a camera for object identification.

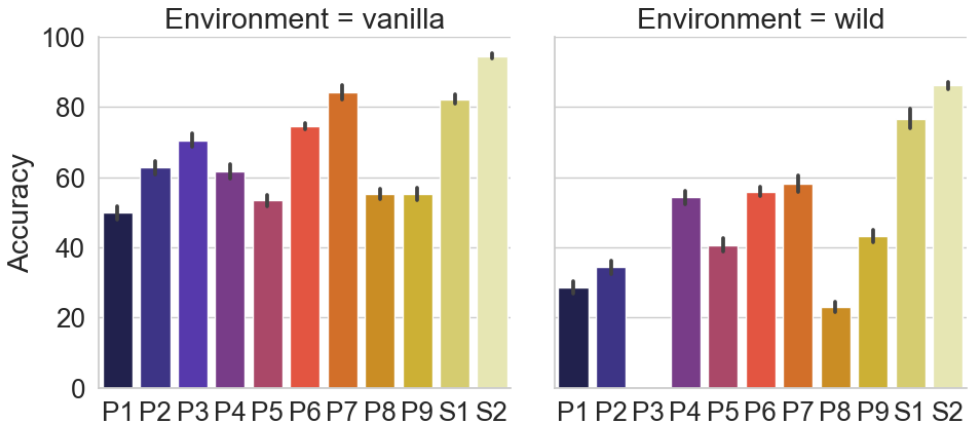


Figure 4.9: Average model accuracy per participant.

Indeed, we observe a negative correlation between our participants' age and the performance of their models ($r = -0.74$, $p < 0.05$) even though this seems to be explained by their photography experience.

Recognition performance varies by photography experience. We observe that participants with experience in photography (P3, P7) tend to achieve higher accuracy than that of those without, with P7's models achieving performance comparable to S1's in vanilla. Yet, this was not the case for all participants with photography experience. For example, P8's models don't outperform by much the models of other participants with no experience. A primary limiting factor seems to be P8's tendency to block the camera view with her finger. Without P8, we observe a correlation between photography experience and model performance in vanilla ($r = 0.8$, $p < 0.05$).

Robustness to new environments is still challenging. Testing on a different environment than the one you trained might perhaps defeat the whole purpose of personalization. However, it also measures robustness. When comparing the vanilla test (same environment with train) to the wild test, we observe that models from S1–S2 achieve similar performance. However, this is not the case for P1–P8. This discrepancy can be explained by many factors. It could be that simply S1–S2 have a better sense of how machine learning works and their visually inspected photos encompass that knowledge. For example, we observed that photos of S1 and S2 in the wild minimized the background with the most salient object viewpoint covering the camera frame. Inaccurate feedback in the cluttered environment for P1–P8 could be another factor. This could explain why the ratio of photos fully including the target object was lower in the wild than in vanilla (Figure 4.10). Last, it could be due to changes in photo variations; variations in vanilla photos seem to be different from those in the wild (Table 4.3).

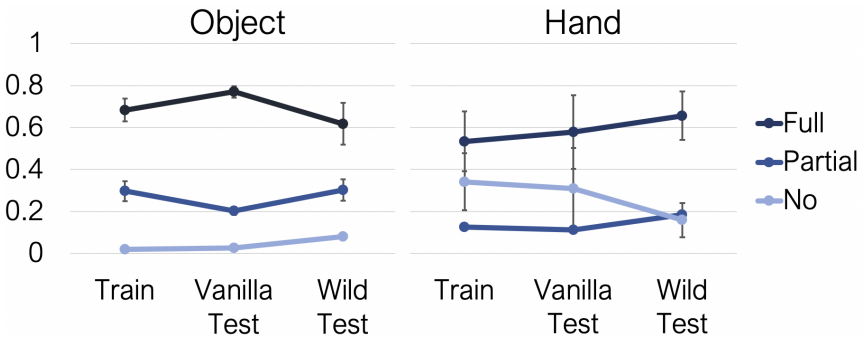


Figure 4.10: Proportion of photos with fully, partially, or not included objects and hands. Errors bars show variance among participants, who were able to fully capture the object in 60–80% of their photos and include their hand in more than 50%.

4.3.2 Qualitative Analysis

Each participant took 375 photos in *train*, 75 in *vanilla test*, and 75 in *wild test*. Participants spent about 4 minutes to train on an object with 25 photos ($153s - 373s$, μ: 258s, σ: 90s).

Photos from the participants were coded using a pre-established coding scheme to analyze common patterns and photo-taking strategies. Beyond counting the presence of the object and hand in the photos (Figure 4.10), we adopted the codes in Hong *et al.* [133] based on the four dimensions that humans generalize across for visual recognition [134]: size, location, viewpoint, and illumination. Specifically, as shown in Table 4.3, we counted how many participants varied the object *size* by zooming in and out; *location* as captured by the difference in the background, and *viewpoint* in terms of object side, perspective such as camera angle, and position in the camera frame. Since we control for lighting conditions, the *illumination* code was not included. Two raters coded the photos independently with an almost perfect agreement (Cohen's kappa=0.84).

Limited number of photos without the object of interest. Figure 4.10 shows the average proportion of photos including the object of interest. We observe that less than 2% of the photos in the train and vanilla test don't include the object of interest. This number is slightly higher

Table 4.3: The number of participants who included variations while taking photos in *train*, *vanilla test*, and *wild test*

	Size	Location	Side	Perspective	Position
(Vanilla) Train	9	1	7	6	9
Vanilla Test	8	1	9	6	9
Wild Test	7	4	5	3	7

(8%) in the wild, where the object localization task and accurate feedback are more challenging.

The inclusion of a hand in photos varies by participants. Figure 4.10 shows the average proportion of photos including a participant's hand. A hand was included in photos when participants either held the objects while taking photos or used their hands as a reference point for photography. Participants were consistent across the three sessions in how they included their hand. P1, P4, P5, P6, P7, and P9 included their hand in more than half of their photos, whereas P2, P3, and P8 in less than 15%. When carefully examining the videos from the latter, we noticed that they used their hand to interact with the object and obtain feedback, and then removed it to use both hands for stabilizing the camera and taking the photo.

Positive attitudes towards teachable object recognizers. At the end of the study, we asked participants about their experience. Overall, they were positive about training and having their own object recognizer (Figure 4.11); all participants agreed that this is feasible, and most of them

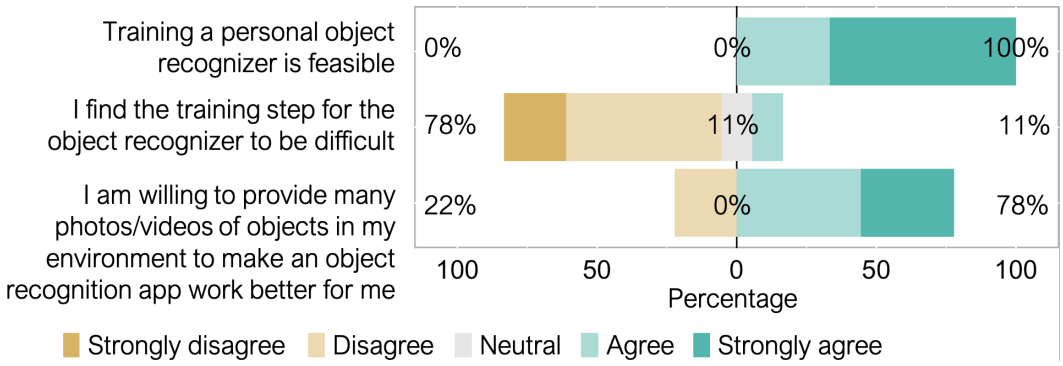


Figure 4.11: Post-study feedback. Most participants thought training is not difficult and they were willing to provide photos to train and improve the performance of an object recognizer.

agreed that this is something that they are willing to do. Only P1, who had no camera experience and was unable to aim well the camera at smaller object stimuli, found training difficult. She said *"I think the small objects are a little bit tough to figure out. There is not much feedback."* Like P1, other participants (P5, P8, and P9) also thought that difficulties arose when there was none or little feedback. P8, who had the tendency to block the camera view with her finger, said *"(I had difficulty with) really small object that we couldn't really get the feedback on."* Even when it was challenging to get feedback for well-centered objects, participants appreciated the spatial information. P5 said *"When you got the decent tone (not the high pitch tone), you knew you were off on that, so at least you got some feedback."*

Focus on the feedback despite it being error-prone. All participants expressed that their attention during the study was on receiving feedback. When asked what did they do when unsure that the feedback was right, they all reported readjusting the position of the phone and the object to get the feedback. Despite having experience with feedback errors during the practice session, P2 and P3 said *"I just had to trust the feedback I had at the last moment"* and *"If (I) didn't get the best feedback, I took the picture based on what feedback I got (before)"*, respectively. Moreover, some participants seemed to utilize the feedback for learning. For example, P4 and P5, who never

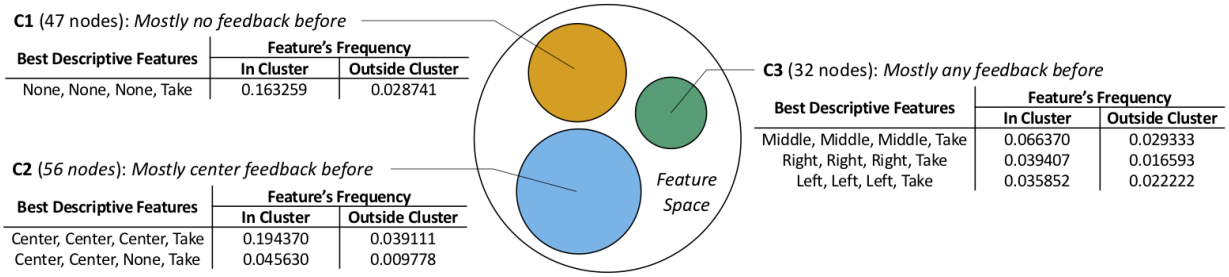


Figure 4.12: By representing the stream of interactions in a (participant, object) pair as a unique node, we find three distinct clusters and their descriptive features that are most responsible for this cluster formation.

took pictures before our study, mentioned *"I found that, in some aspects, it was easier having the object on a flat surface and taking a picture of the object on the flat surface, (but) it depended on what the object was"* and *"To me, (I) have nothing to have keep (my) hand straight, and I intended to go like this, I was using my hand as info (for direction)"*, respectively.

4.3.3 Analysis of Interaction Streams

While the preceding analyses reveal interesting insights, they do not take into account the sequential nature of the feedback experienced by participants. In order to learn richer patterns of camera–object interactions, we adopt some of the unsupervised learning techniques used in Kacorri *et al.* [115] to uncover participant–object clusters based on the interaction streams, which preserve the temporal structure of the data. Feedback data collected during the train mode were used for this analysis, as they characterize the participants' continuous interactions with the feedback on each object.

Specifically, we represent each of 135 participant–object pairs (9 participants $\times$ 15 objects) by the feedback stream received from the server. For each of 25 training photos, we consider any feedback in the preceding 1-second interval, which corresponds to the last three images sent to the server (one every 333ms) for feedback generation. Thus, a stream in our data consists of 25 4-tuples (f_3, f_2, f_1, Take), where f_i could be:

- *None*: No feedback generated (no object location was estimated by our localization model).
- *Center*: The high-tone and haptic feedback generated (an object was estimated to appear at the center, region **E**).
- *Left*: The left-stereophonic tone feedback generated (an object was estimated to appear on

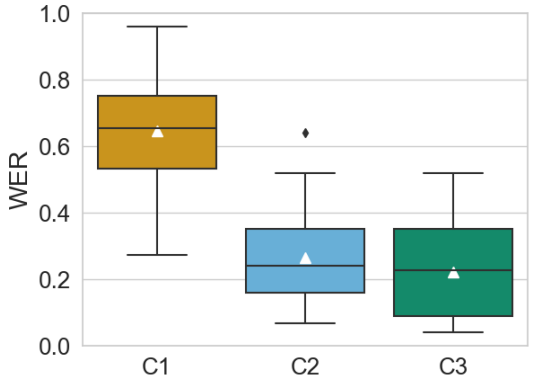
region **A**, **D**, or **G**).

- *Right*: The right-stereophonic tone feedback generated (an object was estimated to appear on region **C**, **F**, or **I**).

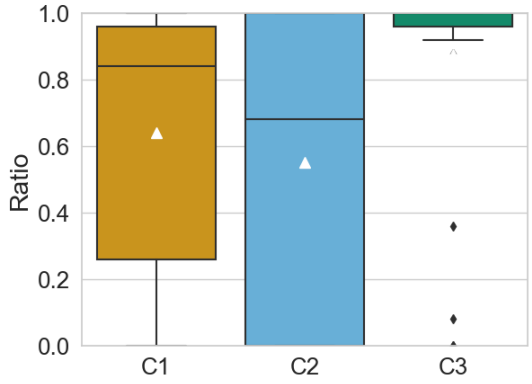
- *Middle*: The mid-stereophonic tone feedback generated (an object was estimated to appear on region **B** or **H**).

We map the pairs to a feature space based on the normalized 4-tuple frequency in each stream, a slight variation from prior work [115, 135, 136]. We then construct a similarity graph by representing each pair with a node and comparing pairs using cosine similarity. Last, we identify clusters of similar pairs by graph partitioning using the Louvain community detection algorithm [137]. As shown in Figure 4.12, to interpret the meaning of the clusters, we isolate the primary features (*i.e.* 4-tuples) that seem responsible for that formation. For each cluster, we build a binary classifier that distinguishes pairs belonging to that cluster from all the other pairs as in prior work [136].

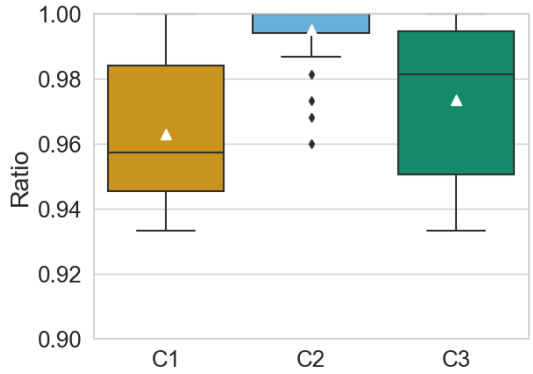
User-object interactions fall under three distinct clusters. The first cluster (**C1**) contains 47 participant-object pairs. The presence of the (None, None, None, Take) tuple, occurring more frequently for the pairs in this cluster that any other pairs, indicates that participants received no feedback immediately before taking these photos. This means that objects were either out of frame or the localization model had a hard time detecting their center. The second cluster (**C2**) contains 57 pairs. In contrast to C1, the higher frequencies of the (Center, Center, Center, Take) and (Center, Center, None, Take) tuples in this cluster indicate that participants held back from taking photos of these objects till they received well-centered feedback. The third cluster (**C3**) contains the other 32 pairs, where participants received some feedback estimating that the object



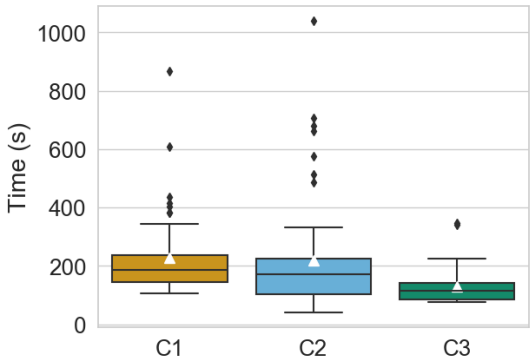
(a) Feedback WER across pairs.



(b) The number of photos including hands.



(c) The number of photos with full object.



(d) Time spent in training.

Figure 4.13: Comparing clusters across different interaction characteristics during training.

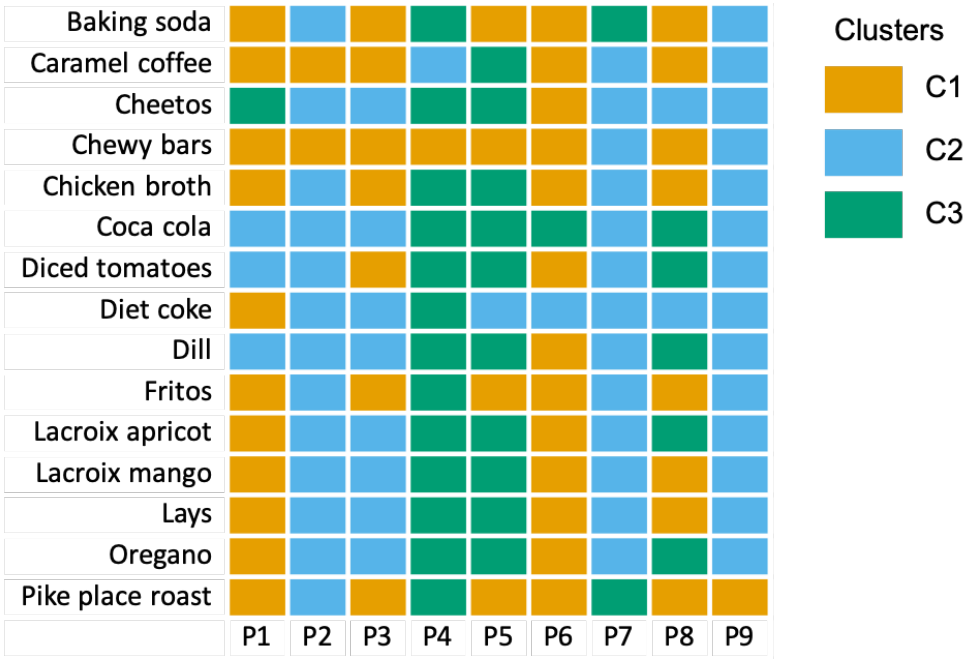


Figure 4.14: Cluster distribution across participant-object pairs.

was in the camera frame, though not well-centered, with higher frequencies for (Middle, Middle, Middle, Take), (Right, Right, Right, Take), and (Left, Left, Left, Take).

Participant-specific & object-specific interactions in clusters. Figure 4.14 shows the distribution of the clusters across participant-object pairs. Reflecting on the fact that our feedback approach relies both on participants' photo-taking strategies and object characteristics, it demonstrates two types of clustering tendencies: participant- and object-wise clustering. Specifically, five participants tend to have consistent interactions captured by a single cluster which were less dependent on the object. For example, P2, P7, and P9 hold off taking the photo till they feel the vibration for well-centered objects and P4 tries to obtain some feedback indicating that the object is in the frame. Only P6, who has some residual sight and often uses other apps for object recognition, tends to take the photo even though there is no feedback. When describing her strategy she said "I put the camera to where I thought the middle of the object was, and then I pulled back and then took a picture". Looking at the object-wise clustering patterns, we observe that participants had difficulty in getting feedback for the *chewy bars* but tend to get well-centered feedback for the diet coke, which was similar to some of the objects the localization model was trained on.

High variance in object localization estimations for feedback. To better contextualize the interaction patterns and participants' feedback in our analysis, we manually annotated all images sent to the server within the 1-second pre-photo window with ground truth object center locations (a total of $10,125$ images; 135 steams $\times$ 3 feedback images $\times$ 25 photos). This allows us to assess the quality of the localization model, which can affect the interactions. Specifically, we calculate the differences between the ground-truth feedback and that received by participants prior to taking the photo using the word error rate (WER) metric. We found a 0.39 WER on

average (0.04 – 0.96 [2], σ: 0.24). As shown in Figure 4.13(a), we also provide pair-level WERs across the clusters.

Different characteristics observed for each cluster. To confirm and further study semantics we associate to each cluster, we characterize the clusters in terms of feedback errors, the ratio of images where hands are present, the ratio of images that included the entire object, and the overall time spent training on an object by the participant (Figure 4.13). We observe that while participants in C1 did not receive feedback right before taking the photo, this is because the localization model often failed to localize the object (WER: μ: 0.65, σ: 0.15). However, they were inconsistent in including their hands in the images, which can help the localization (hand ratio: μ: 0.64, σ: 0.38). Still, many of their training photos fully included the object (μ: 0.96, σ: 0.02) though it took participants a bit longer (μ: 228s, σ: 140s). In contrast, we observe lower WERs for C2 (μ: 0.27, σ: 0.13) and C3 (μ: 0.22, σ: 0.14). For C3, this may be due to the high ratio of images that included the participants' hands (μ: 0.88, σ: 0.30). However, in C2, where the feedback is indicating a well-centered object, we notice a higher variation (μ: 0.55, σ: 0.45). As discussed in the qualitative analysis, this could be explained by some participants including a hand to obtain feedback, then removing it to stabilize the camera and take the photo. One of the most important observations is that *C2 participants that took photos after feedback indicating well-centered objects tend to have a higher ratio of training examples with the object of interest fully included*. Last, we notice that C3 instances tend to spend less time in training (μ: 130s, σ: 66s), with participants taking photos as long as they receive some feedback that the object is within the frame.

[2]Higher WERs were observed on objects such as *chewy bars*, *pike place roast*, and *caramel coffee* that were present in the images but not localized by our model.

4.4 Discussion

We discuss implications and limitations of our findings in the context of blind photography and teachable object recognizers.

4.4.1 Implications

Our study and findings provide evidence for the potential of a real-time feedback approach that can help people with visual impairments include and indicate an object of interest in the camera frame. We see how human- or computer-powered object recognition apps could benefit from the following insights:

- Feedback leveraging hand proximity to the object can help minimize photos that do not include the object of interest even for blind users who have never taken a photo before. Our cluster analysis indicates that more accurate feedback is associated with more photos that include the entire object. This emphasizes the need for more training data to further improve the accuracy of the localization model.
- Even though participants are willing to include their hands to obtain feedback, their hands are often not included in the final photos, which can help to preserve user anonymity when photos are sent to crowdworkers or remote servers. This highlights the need to incorporate such feedback mechanisms on-board the user's device.
- Participants were able to overcome many cases of false negatives (undetected objects). While they did not receive any feedback, they were able to fully or partially include the object in the frame. However, when feedback was received, they tended to trust it, even

though they were exposed to its limitations. This indicates the need for a tighter threshold on the feedback mechanism to reduce false positives.

Our work replicates a previous study by Kacorri *et al.* [4] with an older participant pool and thus contributes to the validation and reliability of their findings. Our results confirm the potential of teachable object recognizers and participants' willingness to personalize their object recognizers by providing training examples. Moreover, we find that:

- There seems to be a negative correlation between participants' age and the performance of their recognizers though this seems to be explained by their photography experience. Thus, it is important to also consider participants with these characteristics when evaluating similar applications.
- One factor limiting recognition accuracy was blocking of camera view with fingers, which can be detected as a special case (similar to detecting no light in the environment).
- Photos taken in simulated or real-world environments with cluttered background remain a challenge and should be included in the evaluation of similar applications.

Last, we demonstrate how to obtain rich insight from data-driven methods on participant interactions with an intelligent feedback mechanism. We see how our clustering and word error rate analysis methods may be adapted to the study of other assistive applications that incorporate real-time feedback.

4.4.2 Limitations

One motivation for teachable object recognizers is that personalized models are required for objects that often don't have a barcode, a readable text, or an accessible database entry but

are unique to users. However, in this study, we use commercial products not meeting these characteristics. While this is a common strategy among researchers [4, 43], which allows for experimental control and replication, it also limits the type of insights and feedback we could gain through real-world scenarios. As we build our technology to a fully working prototype, we will move to real-world deployments.

Another limitation is that our participant characteristics are skewed: all are female and six out of nine are over age 60. However, we think this could be a strength. Studies tend to reach younger people as early adopters. Thus, older adults are often excluded from early stages of technological innovations. We were excited to see our participants training their own intelligent assistive technology, and, for some, this was their first time taking photos. P5, aged 63, said *"I want one today. Because you get a chance of identifying things. And it would be less complex than what I use right now. .. So you could have everything on your shelf that you want to look at, and you could go boom boom, and it would have it in its repertoire."*

When constructing the interaction streams, we limited the feedback to one second prior to each photo in training. While larger windows potentially allow for more insights, it would create an even larger number of images requiring hand-annotations to calculate the performance of the feedback. For this analysis, we annotated more than ten thousand images.

Finally, our analysis did not consider the distribution of confidence scores in the object recognition output; we simply used the label prediction with the highest confidence score (top-1). In a real-world scenario, users may benefit from knowing the model's confidence when it recognizes an object as it may help the users learn more about their recognition model throughout interactions [133]; for example, a recognition model may say "not sure" if the confidence score of its top-1 estimation does not stand out. This is the direction we are currently exploring.

4.5 Summary

This chapter presents a real-time feedback mechanism to help people with visual impairments take better photos of objects with their mobile phones for object identification. By employing convolutional neural networks, our feedback has learned to estimate the center of the object of interest based on its proximity and the pose of the user's hand. The estimated location within the camera frame is communicated to the user through audio-haptic feedback building upon prior work for conveying spatial information in blind photography.

We explore the potential of this feedback mechanism in the context of teachable object recognizers, where people with visual impairments are called to train their object identification application by providing a small number of training examples per object. In a user study with nine participants with visual impairments and two sighted people whose models serve as a baseline, we find that very few photos from the participants do not include the object of interest (2% in the vanilla and 8% in a simulated real-world environment). Moreover, the recognition performance was promising even for those participants who had no prior camera experience.

While many factors can still be explored in improving the quality of the photos (*e.g.* training with better architectures and more data, verbose feedback, and better thresholding), the most important remaining issue is its usability in a real-world setting over longer periods. This chapter has been instrumental for better understanding some of the challenges that users with visual impairments may face when interacting with such teachable interfaces.

Chapter 5: Hand-Primed Object Localization and Recognition on Hand-Held and Head-Worn Cameras[1]

Computer vision holds promise for solving daily challenges that people with visual impairments face; one of them is fine-grained object recognition from an egocentric perspective [5]. Recent breakthroughs in object recognition enable a computer to detect all objects learned from training data. However, without additional indication, it would be challenging for such a model to know which object is of interest to users who are blind. Also, there is no guarantee that, without visual feedback, these users can aim the camera properly to indicate the object of interest in the frame. Consider the photos in Figure 5.1 taken by people with visual impairments who tried identifying objects with the help of a sighted crowd on VizWiz [1] or with their personalized object recognizer [2]. Do all the images contain the object of interest? Do we know about which of the objects is the user inquiring? Does the object show discriminative viewpoints? How can the user tell if the wrong object is being recognized given a cluttered background? These questions highlight the need for non-visual feedback in assistive recognition applications that can guide users to take well-framed photos across different objects of interest. Such guidance is often achieved through a few iterations with sighted help in crowdsourced assistive applications (*e.g.*, BeMyEyes [40] and Aira [24]), but it remains challenging for automated solutions, which may

[1]This chapter presents joint work with Professors Hernisa Kacorri and Abhinav Shrivastava published in IEEE/CVF WACV 2020 [138] and IEEE TPAMI 2021 [139].

Figure 5.1: Examples of egocentric photos taken by people with visual impairments for object recognition in the crowdsourcing app, VizWiz [1], and a personalized fine-grained object recognition model in TEgO [2]. These examples illustrate the need for object-of-interest localization for better camera framing.

lead to recognition errors that are perceptible only through sight.

In response, many non-visual feedback mechanisms have been proposed to help people with visual impairments take better quality photos for identifying objects (*e.g.*, [17, 20]). In particular, hand-based object localization is shown to help blind users better aim their object [31]. While cluttered environments remain challenging, our prior work has demonstrated that hand-object interactions can still help estimate the object of interest [2]. For blind people, hands seem to be a natural and intuitive way to interact with objects when they are taking photos in the context of object recognition [4]. It is further supported by prior evidence on blind people's ability to guide hand orientation and make rapid corrections through proprioception [73].

In this chapter, we explore whether we can leverage this information, carried by hand-object interactions, to benefit object localization and recognition in egocentric vision, which is the ultimate goal of these assistive technologies. We present a computational model that uses *hands* as the contextual information to localize and recognize an object of interest in egocentric static images. In our approach, a pre-trained hand segmentation model provides the hand information as contextual information to guide the detection of the object of interest—similar approaches

have been shown to help a model obtain better context information (*e.g.*, [140, 141]). While bounding box annotations are typically required for object detection (*e.g.*, [6, 7]), our approach relies simply on object center annotations with advantages linked to robustness [103] and cost-effectiveness [142].

We evaluate our approach on three egocentric datasets, GTEA [107], GTEA Gaze+ [108], and TEgO [2]. Focusing on object localization and recognition, our evaluation compares our hand-primed method with other methods that use either only center annotation data or both the center and hand annotations. We find that our approach can contribute to better localization of the object of interest and achieve the recognition performance that is comparable to the state-of-the-art detection model (Faster R-CNN [6]). Note that we have to tweak the Faster R-CNN model to manually select only one detection output closest to the ground-truth object of interest, making this comparison a bit unfair for our method. We further evaluate our method on the photos taken from a blind individual in TEgO. Although it shows that having more annotations (*i.e.*, both hand and object center) helps improve classification, our method still achieves comparable performance to the state-of-the-art model, given fewer annotations (*i.e.*, only object center).

To the best of our knowledge, this is the first work to provide empirical results from different approaches on object-of-interest detection for an assistive application that can empower people with visual impairments. A unique challenge in this task is that gaze information, which is commonly used to infer individuals' areas of interest in egocentric vision (*e.g.*, [108, 111, 143]), is not available for people with visual impairments. However, hand-object interactions can be leveraged to fill this gap.

5.1 Motivation

Prior work in cognitive science has shown that the human visual system tries to use "context" to pay attention only to our interest ([144–146]). The use of the context information has also been discussed from the perspective of computer vision (*e.g.*, [147–149]). Inspired by these observations, researchers in computer vision have introduced various approaches that exploit contextual information (*e.g.*, [150–154]).

In the deep learning era, researchers have also shown that using contextual information, such as semantic segmentation, helps to improve the performance of object detection and recognition (*e.g.*, [140, 141, 155, 156]). FCIS [156] and Mask R-CNN [141] are designed to learn instance segmentation, which is then used for object classification. On the other hand, Shrivastava and Gupta [140] proposed a slightly different approach. In their model, a segmentation module is used to provide contextual feedback in the object detection model (Faster R-CNN [6]); *i.e.*, the segmentation output is appended to the object detection model. However, these existing models are not designed to detect the object(s) of interest to the user—instead, the models are trained to detect all objects that they learn from their training set. Such detection models may not help blind users to know which object among all detected objects is of their interest. To address this issue, we use users' hands as an indication of the object of interest. Our approach takes into account the contextual information (*i.e.*, the hand information) for object recognition in egocentric vision.

Furthermore, recent work explores ways of employing data-driven computer vision methods in assistive technologies. Since such methods often suffer from distributional shifts in deployment ([157–159]), researchers have introduced a teachable interface with which blind users can use to build their personalized system for object recognition (*e.g.*, [3, 4, 160]). In the context

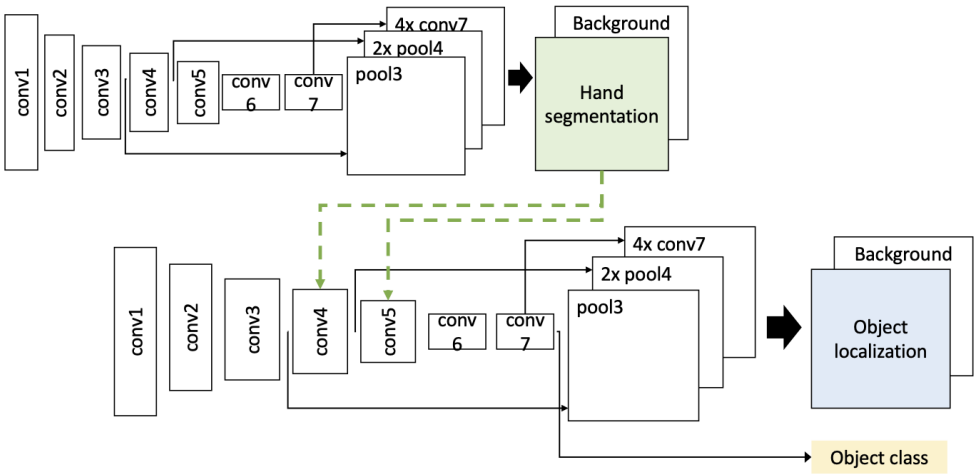


Figure 5.2: The architecture of our hand-primed object recognition model. It consists of two neural networks: one (above) for hand segmentation and another (below) for object localization and recognition. The hand segmentation network is pre-trained, and its output is infused into the other model to help recognize the object of interest being interacted by the egocentric hand. A fully-connected layer for classification is appended to *conv7* of the object localization model. It allows object classification to share the *conv1–conv7* layers of object localization.

of a teachable machine, it is essential to provide good training examples to build a good model. However, providing good training examples is challenging to blind people, especially when they try including objects of interest in the data [17]. Our work tries to explore this issue and evaluates the benefits of our hand-primed method in this context. Also, to enable the users to train their recognizer, it is imperative to reduce data annotation efforts for the training phase. Our method is based on the object center annotation as the point-based annotation is more cost-effective [142] and robust [99] than the bounding box annotation. Our evaluation of object recognition focuses on understanding the feasibility of our method given limited annotation data, compared to the state-of-the-art detection method (*e.g.*, Faster R-CNN [6]) that uses annotated bounding boxes.

5.2 Methods

We introduce a hand-primed object recognition model built upon the prior work on contextual priming and feedback [140]. Our model is designed to use the egocentric hand(s) in a given image to detect an object of interest. Note that this recognition task aims to classify only the object of interest even when there are multiple objects in a scene.

5.2.1 Object Localization

First, we describe our objective of locating an object of interest pertaining to the hand information in egocentric vision. Reinforced by the output from a pre-trained hand segmentation model, our method localizes an object of interest with which an egocentric hand may interact. For both the hand segmentation and object localization models, we employ the FCN-8s architecture [100]. Using the pixel-level classification model, our model tries to estimate which pixels belong to the center area of the object of interest as proposed in [99]—we consider the model proposed by Ma and Kitani [99] as the baseline model in our evaluation of hand-based object localization.

As illustrated in Figure 5.2, the hand segmentation output is appended to two later convolutional layers (*conv4* and *conv5*) of the object localization network to prime the model to use the hand information for object localization[2]. Our approach of infusing the hand segmentation output is inspired by the prior work on object detection primed to segmentation feedback [140]. We design the hand-primed method to incorporate the observations that later convolutional lay-

[2]We tried infusing the hand segmentation to the first five convolutional layers (*conv1–conv5*) for object localization, but observed that this approach did achieve worse performance than the one we introduce in this work. Thus, we describe only one approach in this work.

ers capture the more descriptive features [161] as well as hand-related features in hand-based localization [99].

The hand segmentation model has two output layers to segment hands and a background in an image; *i.e.*, the model performs a per-pixel classification to determine which pixels belong to the hand or the background. For object localization, our model does not estimate a bounding box of the object of interest. Instead, the model infers a center area of the target object. Having two output layers, the object localization model determines which pixels belong to the center area of an object of interest or the non-center area (background). Our model expects egocentric images that are obtained by users taking photography or by extracting salient frames from videos recorded using a wearable camera on the head, the chest, or eyeglasses.

5.2.2 Object Recognition

We further extend our object localization model for object recognition. To classify the object of interest, we propose sharing hard parameters of convolutional layers of the object localization model. A branch for object classification is created in the localization model to enable our model to localize and classify the object of interest pertaining to the hand information from the pre-trained hand model.

As depicted in Figure 5.2, one fully-connected layer, which is to classify the object of interest, is appended to *conv7* of the object localization model. The idea is that sharing all convolutional layers for the localization model would allow the classification task to use features, which are learned for object localization, to recognize the object of interest, and vice versa. In this method, both object classification and localization share most of the convolutional layers

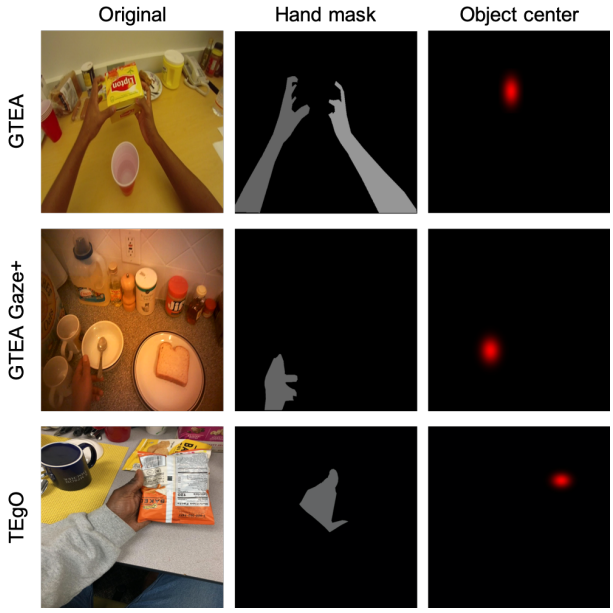


Figure 5.3: Input data for training models: an original image, its hand mask, and its object center annotation.

(*conv1–conv7*) and learn features that help accomplish both of the objectives: estimating the label of the object of interest and its central area.

5.2.3 Implementation Details

To understand the benefits of having the hand and object center information in object localization and classification, respectively, we implemented our hand-primed method in two ways. First, we tried to understand the effects of the hand information solely in object localization. In the first implementation, our model was not trained to classify objects, but to localize an object of interest (*without classification*). The second implementation was based on our recognition model as designed in Section 5.2.2—*i.e.*, we added object classification to the objectives of our model (*with classification*). In the second implementation, our method was trained for both object localization and classification.

5.2.3.1 Without classification

As the object localization output is dependent on the hand segmentation output, we pre-trained the hand segmentation network. Then, we trained only the object localization model. Adam [104] was used to train the hand segmentation and object localization networks. The following are hyper-parameters that we set for training: (hand segmentation) 10,000 training steps, 0.00001 learning rate, 16 batch size, and 10^{-9} Adam's epsilon; (object localization) 20,000 training steps, 0.00001 learning rate, 8 batch size, and 10^{-9} Adam's epsilon. In both of the models, we initialized the weights of the first five convolutional layers (*conv1–conv5*) with those of the VGG-16 [162] pre-trained on ImageNet [163].

Pre-training: Using four egocentric datasets including EgoHands [110], GTEA [107], GTEA Gaze+ [108], and TEgO [2], we first pre-trained a hand segmentation model. Provided with a set of original images and those hand masks as shown in Figure 5.3, the hand segmentation model was trained to learn hand features from the input data. The pixel-wise loss was calculated with the cross-entropy function. The loss function was the same as Equation 5.1, but there were two classes for each pixel estimation—*hand* and *non-hand*.

Training: To train our object localization model, a set of original images and those object center annotation data were used; Figure 5.3 shows the examples. The center of an object of interest in each image was annotated with a Gaussian heatmap blob. The object localization model was taught to estimate the center area of a target object interacted with the hand. GTEA, GTEA Gaze+, and TEgO were used to train and test the object localization model. We did not use the EgoHands dataset for object localization as this dataset mostly contains hand movements in board games, such as chess, card games, and Jenga. As our localization model predicts a class

per pixel, we used the cross-entropy loss function for training. The pixel-wise loss is repeatedly calculated and summed over all output pixels, and then averaged as defined in Eq. 5.1—I denotes the set of output pixels, and C denotes the number of classes per pixel.

$$L_{\text{local}} = -\frac{1}{|I|} \sum_{i \in I} \sum_{c \in C} y_{i,c} \log(\hat{y}_{i,c}) \tag{5.1}$$

Inference: At test time for object localization, we only considered per-pixel classification outputs that had the confidence score higher than 0.5 in the output layer. When there was more than one cluster of the estimated center area, the biggest cluster was chosen to be the estimated center area.

5.2.3.2 With classification

We implemented the additional branch for object classification on our hand-primed object localization model. As before, we initialized the weights of the first five convolutional layers (*conv1–conv5*) of the model and used Adam [104] to train our model. The following are hyper-parameters used in training: 40,000 training steps, 0.00001 learning rate, 4 batch size, and 10^{-9} Adam's epsilon.

Training: To train the model with classification, we input three different data—an original image, an object center annotation, and a target object's label. Equation 5.2 shows the cross-entropy loss function for the classification task—O denotes the number of object labels. In the end, to train our object recognition model, we use L_{recog} that combines L_{local} and L_{cls} as defined

in Equation 5.3.

$$L_{\text{cls}} = -\sum_{o \in O} y_o \log(\hat{y}_o) \tag{5.2}$$

$$L_{\text{recog}} = L_{\text{local}} + L_{\text{cls}} \tag{5.3}$$

Inference: Similar to our object localization evaluation, a threshold of 0.5 was used to obtain an object localization output (per-pixel classification). As described earlier, we chose the biggest cluster to be the estimated center area of an object of interest when there was more than one cluster in the object localization output,. For the classification task, we considered the top-1 class, which has the highest confidence score, as a predicted object label.

5.2.4 Datasets

All the datasets (GTEA, GTEA Gaze+, TEgO) already provide hand segmentation data for their original images. However, only the TEgO dataset provides the annotation data for object location; they used a Gaussian heatmap blob to indicate the center of an object of interest, which was also adopted for egocentric activity recognition [99]. For the other datasets (GTEA and GTEA Gaze+), we manually annotated these datasets for the object localization; we shared these annotation data on the TEgO website. Note that the heatmap blob was drawn on the object's center point, and its blob size was the same throughout each dataset. The following is the number of images used from each dataset: GTEA (663), GTEA Gaze+ (1,115), and TEgO (5,758). Note that we excluded images without hands from the TEgO dataset in our evaluation. We randomly selected samples from each dataset and applied the following dataset splitting to each dataset:

80% for training, 10% for validation, and 10% for testing.

The GTEA datasets include images collected only by sighted people, but the TEgO dataset contains data from a blind individual. The evaluation with this TEgO data would gauge each localization model's compatibility with assistive technologies for people with visual impairments, such as blind photography or object recognition. In the TEgO dataset, 19 objects were used, and there are three different geometrical types—boxes, tubes, and cylinders. Two cereal boxes and two snack bags are in the box type while the tube type includes one hand lotion and one sunscreen. The rest belongs to the cylinder type, but has different shapes—two small spice jars, two food cans, two soda cans, three soda bottles, two medicine/nutrient bottles, and two hand-soap/hand-sanitizer dispensers.

5.2.5 Comparison Models

We first implemented four different models to evaluate hand segmentation and object localization. All the comparison models were also built upon the FCN-8s model. As in our method, Adam optimizer was used for training. For object classification evaluation, the branch for object classification was implemented on all the comparison models. We optimized the hyperparameters and ensured that the number of training images each model saw was the same—*i.e.*, the product of the training steps and the batch size is identical throughout all the models. Two versions of Faster R-CNN [6] were implemented as additional comparison models—thus, there are a total of six comparison models to evaluate the classification performance. For the Faster R-CNN models, we used different hyperparameters since these models were built on a different architecture.

NoHandData: This model was trained[3] for object localization without explicitly being

[3] 10,000 training steps, 0.00001 learning rate, 16 batch size.

trained on hand segmentation data; *i.e.*, the object center annotation was only used for the experiment without classification, and the object label was added for the experiment with classification.

Finetune: This is the baseline model that uses hand features for object localization, which was introduced by Ma and Kitani [99]. To implement this model for the evaluation without classification, we fine-tuned[3] our hand segmentation model to the object localization problem for each dataset. For the experiment with classification, the hand model was fine-tuned to both the object localization and classification tasks. During the fine-tuning, we froze the weights of the first five convolutional layers (*conv1–conv5*) as in the model proposed by the prior work; they observed that hand features were captured in the conv5 layer.

MultiClass/MultiClass-2x: In this model, learning the hand and the object center was considered as a multi-class problem. This multi-class model estimates the class of each pixel; *i.e.*, it determines to which class among *background*, *hand*, and *object center area* each pixel belongs. In addition, we implemented another multi-class model, *MultiClass-2x*, in which we set twice more weight on the loss in object localization. For both of the models, the weights were also initialized with those of our object hand segmentation model before training[3]. In the experiment with classification, the models with the classification branch were trained with the objects' labels.

MultiTask/MultiTask-2x: Considering hand segmentation and object localization as two different tasks in one model, we employed the hard parameter sharing for this multi-task learning [164]. This model shares the first five convolutional layers (*conv1–conv5*) and has two separate branches—one branch for the hand segmentation and another for the object localization. This model is called *MultiTask*. Moreover, as in the MultiClass models, we also created another multi-task model in which the loss in object localization model was computed to be twice more

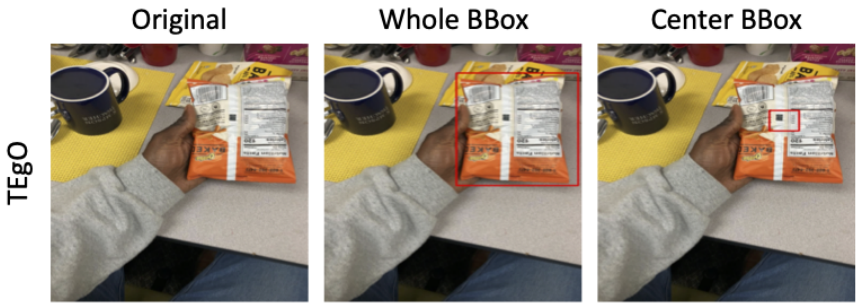


Figure 5.4: Input data for two different versions of Faster R-CNN: an original image, a whole object bounding box, and an object center bounding box.

important than the loss in hand segmentation. We call this model *MultiTask-2x*. Before training[4] these models, we also initialized the weights of the hand segmentation part of these models with those of our hand segmentation model; *i.e.*, the weights of the shared convolutional layers and the separate layers for the hand segmentation part were initialized with those of our hand segmentation model. As for the classification evaluation, the third branch for classification was enabled on these models.

Faster R-CNN (with whole bbox/with center bbox): We also used Faster R-CNN [6] to evaluate our approach, which is based on the object center annotation, compared to the bounding-box approach in object detection. Two versions of Faster R-CNN were implemented—one trained with annotated bounding boxes surrounding the whole shape of the objects of interest, and another trained with bounding boxes of the objects' center. The object center bounding box is generated from the Gaussian heatmap blob indicating the object center. Figure 5.4 shows the examples of input data. Similar to the other models, we used VGG-16 as the backbone network for Faster R-CNN and initialized it with the pre-trained weights. Both of these detection models were trained on TEgO with stochastic gradient descent with 0.9 momentum parameter, 40,000 training steps, 0.001 learning rate, and 256 batch size. The learning rate decreased to 0.0001 after

[4] 20,000 training steps, 0.00001 learning rate, 8 batch size.

30,000 training steps. Although we did not try fine-tuning these hyper-parameters to the TEgO dataset, we found this setting, which was used for other image datasets, such as COCO [165], suitable for TEgO. In addition, we manually created the whole-object bounding box annotation data for TEgO as this dataset does not provide it. Note that we use the TEgO dataset to compare our object-center-annotated approach with this bounding-box-based approach in object localization and classification. Furthermore, the state-of-the-art models are not designed to localize only the object of interest—*i.e.*, they can detect more than one object. In our evaluation, therefore, we manually selected and used only one bounding box estimation, which center is the closest to the ground-truth object center.

5.3 Results

Using egocentric hand–object datasets, we compare our hand-priming approach with several other approaches of using the hand information to localize and recognize the object of interest. Our evaluation first focuses on understanding how the hand information can guide the detection of the object of interest. Then, we explore whether object classification would benefit from having explicit features for the hand segmentation.

As the localization outputs and these ground-truths are represented as an area, $mIoU$ (mean intersection over union) and the standard COCO metrics, including AP (averaged over IoU thresholds), AP_{50}, and AP_{75}, are used to report the performance of each model. AP is used to capture false-positive predictions [166], which are a critical factor in object localization. It measures the capability of a model to be employed in assistive technologies for people with visual impairments, such as blind photography, since users of such systems are prone to

Table 5.1: Quantitative analysis of the hand segmentation models. The mean intersection of union ($mIoU$) and the COCO standard metrics (AP, AP_{50}, AP_{75}) are used.

Dataset	Model	$mIoU$	AP	AP_{50}	AP_{75}
GTEA	NoHandData	N/A	N/A	N/A	N/A
	Finetune	0.90	0.85	**1.0**	**0.98**
	MultiClass-2x	0.77	0.62	**1.0**	0.64
	MultiTask-2x	**0.92**	**0.87**	**1.0**	**0.98**
	Ours	0.90	0.85	**1.0**	**0.98**
GTEA Gaze+	NoHandData	N/A	N/A	N/A	N/A
	Finetune	**0.91**	0.85	0.98	**0.95**
	MultiClass-2x	0.83	0.71	0.96	0.80
	MultiTask-2x	**0.91**	**0.87**	**1.0**	0.94
	Ours	**0.91**	0.85	0.98	**0.95**
TEgO	NoHandData	N/A	N/A	N/A	N/A
	Finetune	**0.92**	**0.91**	**0.99**	**0.98**
	MultiClass-2x	0.89	0.84	0.98	0.95
	MultiTask-2x	**0.92**	**0.91**	0.98	0.97
	Ours	**0.92**	**0.91**	**0.99**	**0.98**

false-positives [31]. For object classification evaluation, we use common metrics, such as $F1$, $precision$, and $recall$. Using these metrics, we report the models' performances in a way that 1.0 indicates the highest performance—*i.e.*, the closer to 1.0, the better performance.

5.3.1 Without Classification

5.3.1.1 Hand segmentation

We first evaluate the hand segmentation performance of the models as our work focuses on this information to localize an object of interest in egocentric vision. A threshold of 0.5 was used to determine the per-pixel classification on the hand segmentation output layer. We measured the hand segmentation performance of *Finetune* before fine-tuning it to object localization; thus, two models (*Finetune* and *Ours*) ended up having the same hand segmentation model. Note that the hand segmentation part for the multi-class (*MultiClass-2x*) and multi-task (*MultiTask-2x*) models

was initialized with the weights of our hand segmentation model and then trained on each dataset.

Quantitative analysis: We report the quantitative analysis of the hand segmentation performance of each model in Table 5.1. Regarding the multi-class and multi-task models, it shows only *MultiClass-2x* and *MultiTask-2x* as these models overall outperform their naive models, *MultiClass* and *MultiTask*, respectively. Note that *NoHandData* was not evaluated for hand segmentation since this model was trained only for object localization. As described in the table, our hand segmentation model and the multi-task models show equal to or higher than 0.9 in mIoU and 0.85 in AP in all of the three datasets. On the other hand, the performance of the multi-class model is not competitive to the other models, including our hand segmentation model. Since the multi-class setting is unable to develop independent features for the hand and object center, respectively, we suspect that features gained from the multi-class learning might not be as effective to perform fine-grained hand segmentation as those in the other settings, such as multi-task learning.

Qualitative analysis: The hand segmentation outputs of all the models (except for *NoHandData*) are visualized with the green color in Figure 5.5. As in the quantitative analysis, our visual inspection informs that all the models segment hand(s) from testing images, but the multi-class models lack detailed segmentation. In general, the multi-class models segment a large part of the hand(s) from the testing images, but sometimes are unable to provide fine-grained hand segmentation. In the figure, we observe that the multi-class models are unable to segment fingers from the testing examples from GTEA and GTEA Gaze+.

Figure 5.5: Hand segmentation outputs of the comparison models and our approach. The hand segmentation is overlaid with the green color. All the models except for the multi-class models appropriately segmented whole hand(s) from the testing examples; the multi-class models failed to segment fingers from the examples from GTEA and GTEA Gaze+.

5.3.1.2 Object localization

As in the hand segmentation analysis, we evaluate the object localization performance, quantitatively and qualitatively. A threshold of 0.5 was used to determine whether per-pixel estimations belong to the center area of an object.

Quantitative analysis: Table 5.2 shows the object localization performances of our hand-primed model and the comparison models. In the GTEA and TEgO datasets, our model outperforms the comparison models in almost all of the metrics. In the GTEA Gaze+ dataset, our model show the best performance in AP_{75} and comparative performances to those of *MultiClass-2x* in the other metrics, such as $mIoU$, AP, and AP_{50}.

The results of our model indicate that modulation of hand features in higher layers helps capture the contextual relationship between hands and the object of interest. Compared with

Table 5.2: Quantitative analysis of the object localization models. The mean intersection of union ($mIoU$) and the COCO standard metrics (AP, AP_{50}, AP_{75}) are used.

Dataset	Model	$mIoU$	AP	AP_{50}	AP_{75}
GTEA	NoHandData	0.69	0.46	0.79	0.50
	Finetune	0.68	0.45	0.79	0.41
	MultiClass-2x	0.66	0.45	0.76	0.45
	MultiTask-2x	0.67	0.47	0.77	0.48
	Ours	**0.73**	**0.51**	**0.81**	**0.52**
GTEA Gaze+	NoHandData	0.49	0.26	0.53	0.26
	Finetune	0.43	0.26	0.42	0.21
	MultiClass-2x	**0.60**	**0.38**	**0.77**	**0.32**
	MultiTask-2x	0.52	0.33	0.69	0.29
	Ours	0.55	0.33	0.74	0.24
TEgO	NoHandData	0.69	0.49	0.87	0.48
	Finetune	0.70	0.50	0.86	0.51
	MultiClass-2x	0.70	0.46	0.90	0.41
	MultiTask-2x	0.72	0.51	**0.93**	0.51
	Ours	**0.74**	**0.54**	0.92	**0.57**

the baseline model (*Finetune*), our approach achieved the performance improvement by 13% (from 0.45 to 0.51 AP), 27% (from 0.26 to 0.33 AP), and 8% (from 0.50 to 0.54 AP) in GTEA, GTEA Gaze+, and TEgo datasets, respectively. This result demonstrates that the explicit use (infusing) of the hand segmentation in object localization is more helpful for localizing an object of interest than fine-tuning the fixed features of the hand model to a different problem, object localization. Furthermore, in comparison to *NoHandData* that did not have explicit training for hand information, our model shows better object localization performance in all the three datasets. Based on these observations, it seems that having explicit hand features in the last convolutional layers (*e.g.*, *conv4* and *conv5*) helps to prime the object localization model to focus on necessary features, such as hand pose and location.

Qualitative analysis: Figure 5.6 shows that our hand-primed model well predicts the center area of an object associated with hand(s). As described in the quantitative analysis, *Ours*

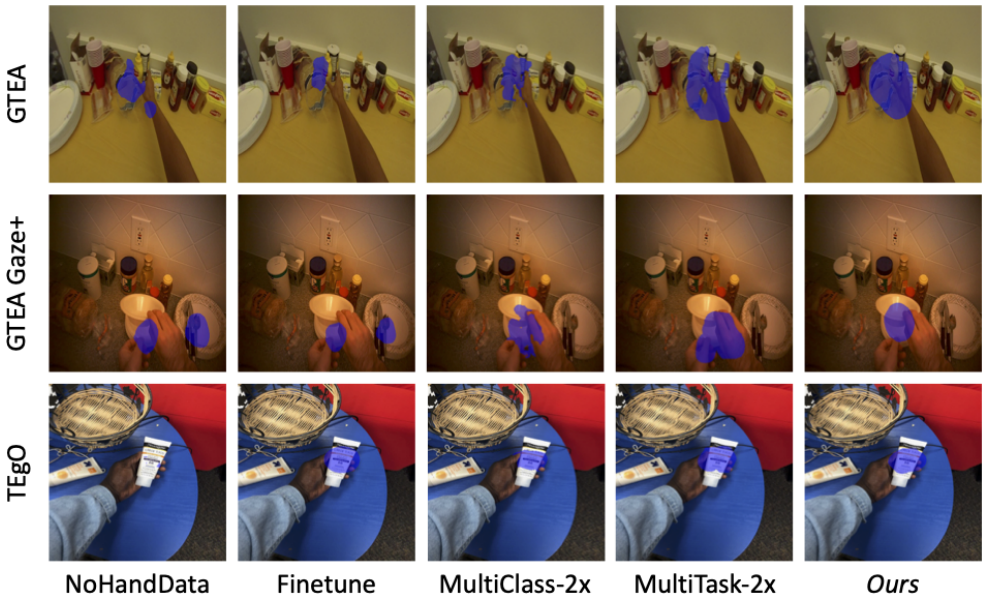


Figure 5.6: Object localization outputs of the comparison models our approach. The object localization output is overlaid with the blue color on the testing examples; the threshold for the localization output was set to 0.5. In comparisons with *NoHandData*, our method (*Ours*) shows that the hand information helps localize the target object.

generate more precise localization outputs than the comparison models in general; the testing examples depict that *Ours* tend to include the center area than the comparison models and is able to estimate only one center area, correctly—*e.g.*, some of the comparison models, such as *NoHandData* and *Finetune*, sometimes estimate more than one center area as shown in the GTEA Gaze+ example. Moreover, our visual inspection reveals that *Ours* estimate a round shape of the center area, which is closer to the ground truth, than do the other models. In particular, comparing the output of our method with that of *NoHandData*, we find further evidence that the hand information is essential not only to localize an object of interest but also to decide which object would be of interest when multiple objects appear in the scenes.

Failure cases: Figure 5.7 shows failures cases of our model in object localization. It can still be confused with two or more objects when the hand–object interaction is not so obvious. For example, the left example in the figure shows that our model is confused with the two objects

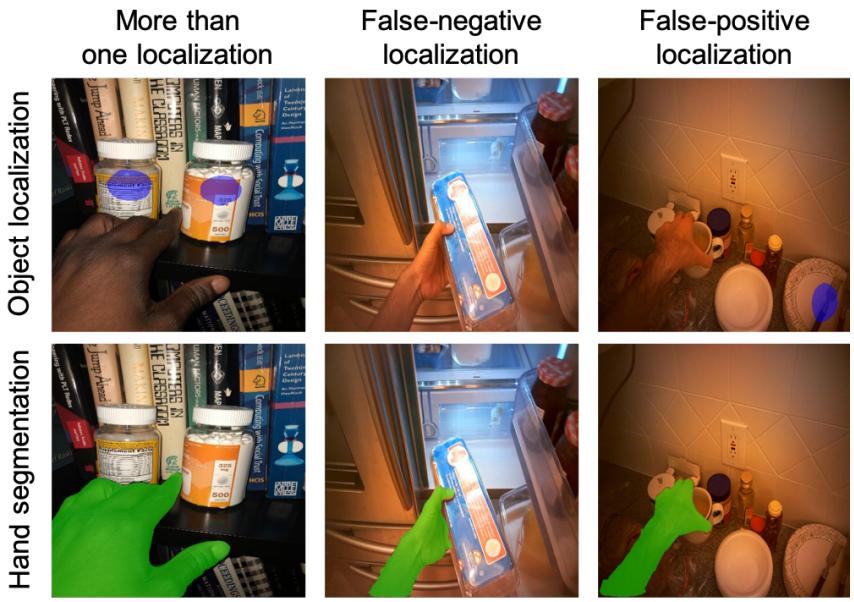


Figure 5.7: Failure cases of our method despite the appropriate hand segmentation outputs on these testing examples.

in proximity of the hand. This confusion seems to be caused by an indiscernible interaction between the hand and the object in the example. In addition, the middle example shows the false-negative localization of our model. Although our hand model successfully segments the hand, the localization model is unable to detect the object being held by the hand in the example. Furthermore, the false-positive output of our model is presented on the right of the figure. Despite the appropriate hand segmentation, our model localized a wrong object, which might be caused by the object occlusion by the hand. Further investigations are necessary for this.

More in TEgO: Data from blind people may not have some implicit information about the object of interest—*e.g.*, placement of the target object at the center of the camera frame. Hence, in data from blind people, hands may be a reliable indication of a target object. As only TEgO includes data collected by a blind individual, we use this dataset to further evaluate all the methods after dividing the TEgO dataset into two subsets—one set including data from the sighted and the

other set including data from the blind. The state-of-the-art detection model (Faster R-CNN [6]), which uses a bounding box for object detection, was employed for this evaluation as well. In Table 5.3, we report the performance of each model on data from sighted and blind people in the TEgO dataset, respectively. The table shows that the Faster R-CNN model trained with whole bounding boxes (*i.e.*, the bounding box surrounding a whole object) shows the best performance in object detection in terms of $mIOU$ and AP. It shows that the cost of annotating a whole object with the bounding box contributes to the accurate detection, although the current metrics ($mIOU$ and AP) may not be appropriate to compare our center-heatmap approaches with the whole-bounding-box approach—almost all of the whole bounding boxes are much bigger than the area of the object center annotations. Hence, we instead use the other version of Faster R-CNN, which was trained with center bounding boxes (*i.e.*, the bounding box including an object's center area), in our comparison. The table shows that *Ours* overall outperforms the other methods on TEgO (sighted) but also on TEgO (blind) in object localization. In particular, compared to the baseline model (*Finetune*), *Ours* achieves the performance improvement, on average, by 8% (0.51 to 0.55 AP) and 5% (0.59 to 0.62 AP) on the TEgO (blind) and TEgO (sighted) datasets, respectively. In addition, when comparing *Ours* with *Faster R-CNN (w/ center bbox)*, we observe that our model improves the localization performance by 45% (0.38 to 0.55 AP) and 77% (0.35 to 0.62 AP) on the TEgO (blind) and the TEgO (sighted), respectively.

5.3.2 With Classification

Among the three egocentric datasets (GTEA [107], GTEA Gaze+ [108], TEgO [2]) used for our object localization evaluation, only the TEgO dataset provides ground-truth labels for

Table 5.3: Object localization performances of the methods on a different set of the TEgO data. For this analysis, the data from the blind person are separated from the data from the sighted person. The bold indicates the best among the models except for *Faster R-CNN (w/ whole bbox)*.

Dataset	Model	$mIoU$	AP	AP_{50}	AP_{75}
TEgO (blind)	*Faster R-CNN (w/ whole bbox)*	*0.92*	*0.88*	*0.99*	*0.98*
	Faster R-CNN (w/ center bbox)	0.59	0.38	0.74	0.34
	NoHandData	0.71	0.51	0.89	0.54
	Finetune	0.71	0.51	0.90	0.53
	MultiClass-2x	**0.75**	**0.56**	**0.91**	0.59
	MultiTask-2x	0.71	0.48	0.88	0.49
	Ours	**0.75**	0.55	**0.91**	**0.61**
TEgO (sighted)	*Faster R-CNN (w/ whole bbox)*	*0.93*	*0.92*	*1.0*	*1.0*
	Faster R-CNN (w/ center bbox)	0.54	0.35	0.65	0.30
	NoHandData	0.74	0.58	0.92	0.64
	Finetune	0.74	0.59	0.96	0.61
	MultiClass-2x	0.73	0.55	0.93	0.55
	MultiTask-2x	0.74	0.56	**0.98**	0.55
	Ours	**0.76**	**0.62**	0.95	**0.65**

objects of interest. Hence, we used the TEgO dataset for our object recognition evaluation. The configuration of splitting the dataset for training, validating, and testing object recognition models was the same as the one for our object localization—samples were randomly selected and split into three subsets, training (80%), validation (10%), and testing (10%). The TEgO dataset provides 19 different object labels for a classification task. Some of the objects fall under the same category and thus need fine-grained classification. There are three categories in terms of objects' geometrical types—box, tube, and cylinder. For detail, please refer to Section 5.2.4. In addition to the Faster R-CNN models, we added another baseline model trained only for the object classification task—*i.e.*, we did not train this model explicitly for object localization and hand segmentation. We refer to this model as *NoHandData (without center blob)*. Comparing it with *NoHandData (with center blob)*, whcih was trained with the center annotation as well, we investigated the effects of object localization in the classification task.

Table 5.4: Quantitative analysis of object classification on the TEgO dataset. F1-score, precision, and recall (*macro average*) are used.

Model	$F1$	$Precision$	$Recall$
Faster R-CNN (w/ whole bbox)	*0.98*	*0.98*	*0.98*
Faster R-CNN (w/ center bbox)	0.89	0.94	0.85
NoHandData (w/o center blob)	0.95	0.96	0.95
NoHandData (w/ center blob)	0.98	**0.99**	0.98
Finetune	0.97	0.97	0.97
MultiClass-2x	**0.99**	**0.99**	**0.99**
MultiTask-2x	**0.99**	**0.99**	**0.99**
Ours	0.98	0.98	0.98

Quantitative analysis: As presented in Table 5.4, the NoHandData model with object localization training achieves higher performance in object classification than that without the localization learning. This comparison result indicates that object localization indeed works as contextual priming for the classification task and reinforces its learning. Furthermore, the table shows that our hand-priming approach achieves comparable performance in object classification to the state-of-the-art approach (*i.e.*, *Faster R-CNN (w/ whole bbox)*) that benefits from annotated bounding boxes. Note that we had to make a favorable modification to the Faster R-CNN models for our evaluation since these models can result in more than one object detection—*i.e.*, among multiple estimations from these models, we only chose the one closest to the ground-truth object.

The best performances of *MultiClass-2x* and *MultiTask-2x* indeed demonstrate that using both the object center and hand annotations helps these models to learn the object of interest more accurately. Such models, however, require to have more data available for training. It contradicts our reasoning of choosing the object center annotation over the whole object bounding box annotation, which is to minimize the data annotation cost in assistive technologies. Although our hand-primed model does not outperform all the other models, our model using less expensive annotation data (*i.e.*, the object center) is able to achieve the performance of the state-of-the-art

Figure 5.8: Cases when the classification task helps object localization. The previous models without classification fail to locate the target object, but become able to localize it after being trained for classification.

algorithm in object classification.

Qualitative analysis: We suspect that the classification task may have guided a model to learn distinguishable features that are helpful for not only object classification but also object localization. To understand the effects of the classification task in object localization, we visually inspect and compare localization outputs of the models with their previous models, which are the ones without the classification training. As depicted in Figure 5.8, adding the object classification task to the training process helps improve the localization task. The examples demonstrate that the models become able to localize the object of interest after being trained with the classification task.

On the other hand, Figure 5.9 shows that there are still a few cases in which the models are unable to localize the object of interest. As shown in the examples from the Finetune and MultiTask-2x models, some of the models become unable to estimate the object center area after being trained for both object localization and classification. Further research would be necessary to understand the failures cases.

Figure 5.9: Examples show that models sometimes become confused or unable to locate an object of interest when trained for both object localization and classification (*With Classification*).

Table 5.5: Object classification performances on the blind individual's data in TEgO. The methods were trained and tested on TEgO (blind).

Model	$F1$	$Precision$	$Recall$
Faster R-CNN (w/ whole bbox)	*0.98*	*0.98*	*0.98*
Faster R-CNN (w/ center bbox)	0.86	0.93	0.84
NoHandData (w/o center blob)	0.87	0.90	0.87
NoHandData (w/ center blob)	**0.99**	**0.99**	**0.99**
Finetune	0.96	0.96	0.97
MultiClass-2x	**0.99**	**0.99**	**0.99**
MultiTask-2x	0.97	0.98	0.98
Ours	0.98	0.98	0.98

More in TEgO: Our objective is to employ object localization and recognition in assistive technologies for people with visual impairments. To understand how each model could work in the target domain, we separate the blind individual's data from the TEgO dataset and measure the classification performance of each model on it. Table 5.5 shows the classification performance of each model on the data from the blind. As in our previous classification evaluation, the best performance of the multi-class model (*MultiClass-2x*) suggests that it would be more helpful to provide more data for a deep learning model so that it can learn from the plentiful data to create distinguishable features for object classification. However, our hand-priming approach

still achieves comparable performance to that of the state-of-the-art model without requiring more data for the training.

Moreover, we observe that there is the performance gap between two versions of NoHandData—*NoHandData (w/o center blob)* trained without object center annotation, and *NoHandData (w/ center blob)* trained with the center annotation. This performance gap indicates that an additional cue to the object of interest is necessary to recognize the object of interest, especially in images from the blind population. Seeing the high classification performance of the NoHandData model trained with the center annotation, we suspect that object classification can benefit from more explicit information regarding the object of interest.

5.4 Discussion

5.4.1 Implications

In an assistive system for blind users, it may be impractical to assume that an object recognition model can recognize all objects blind users may want to recognize, especially since fine-grained classification is often more critical for this population. Instead, it might be more practical to allow blind users to personalize the model by fine-tuning it with photos and labels of their objects. The accessibility research community has been exploring such interactive machine learning approaches, called teachable object recognizers for the blind (*e.g.*, [4, 160]). In such teachable assistive technologies, it is essential to help the users provide good training examples (*i.e.*, photos including the object of interest) and require no additional annotation other than object labels for training, making our approach well suited to this task.

In our evaluation, we learn that more annotations (*e.g.*, the object center as well as the hand

annotations) can enhance the performance of object classification even when photos are captured by blind individuals. In addition, the simple annotation of the object center helps classify the object of interest, although it is not as helpful as the hand information in object localization. In the TEgO (blind) dataset, our hand-primed model, which requires only the object center annotation for training, not only shows the best performance in object localization but also achieves as high performance in object classification as the Faster R-CNN model [6], which is trained with bounding boxes of whole objects. This finding is promising in that our model requires relatively low-cost data to build an object localization model and then further extend it for object classification. Observing the performance improvement from the NoHandData models in classification, we see that the localization output of our hand-primed method can also be used to reinforce object classification, especially during training—it might reduce data annotation cost in teachable object recognition.

There is prior work that tries to understand hand-object manipulation and estimate object attributes from hand grasp types [118]. Since our hand-primed model encoded hand information, such as hand pose and location, our initial expectation was that such explicitly-encoded hand features could help to classify an object as well. However, there are some discussion points regarding the lower performance of our approach than some of the comparison models in object classification (Table 5.4). First, the hand in the TEgO dataset sometimes is either placed in proximity of the object of interest or pointing the object; *i.e.*, the hand is not explicitly grasping the object in some of the data. In addition, the TEgO dataset includes many similar objects that are under the same category, such as bottles of coca cola, diet coke, and sprite, which lead to similar hand grasp types. In such fine-grained classification, hand grasp types may not be as much useful information.

5.4.2 Limitations

A larger dataset from the blind population could contribute to more generalizable evaluation results. Research in assistive systems powered by state-of-the-art computer vision is hindered by the scarcity of available population data ([1, 167]). This is partly because focusing on a small portion of the population leads to smaller datasets. Combined with privacy concerns, disparate characteristics even within a given disability, and need for domain knowledge or in-house annotations can lead to unique challenges for collecting large accessibility datasets [3]. Our research group is committed to responsible data sharing, contributing with the first accessibility data surfacing repository called IncluSet[5] [168], making accessibility datasets searchable by mainstream engines such as Google Dataset Search[6]. Aligned with disability community values, we will continue exploring the performance of our proposed methods with more population data; we are currently in the process of annotating data from nine blind participants interacting with a teachable object recognizer in a lab setting [31].

As observed in the GTEA and GTEA Gaze+ datasets, egocentric data collected with wearable cameras can include both left and right hands and two different objects being manipulated by the hand, respectively. In such cases, therefore, localizing an object of interest for each hand would be necessary to understand egocentric activities more accurately. In addition, prior work shows that video data has temporal information that may help understand hand–object manipulations (*e.g.*, [109,169,170]). We also think that our method would benefit from a more suitable loss function. As our model estimates the object's center area, a loss function that can compute losses between the ground-truth area and the estimated area could be more appropriate and helpful to

[5]https://incluset.com/
[6]https://datasetsearch.research.google.com/

our approach.

5.5 Summary

In this chapter, we introduced a hand-primed model that localizes and recognizes an object of interest by incorporating hand information. Our evaluation showed that the explicit infusion of the hand information to object localization helps locate the object of interest in egocentric vision, and this is evident both in data from sighted and blind individuals. With lower-cost annotation data for object localization (*i.e.*, object centers), our hand-primed model achieved as high performance in object classification as the Faster R-CNN model that relies on bounding-box annotations for the objects of interest. Although more annotation data can help improve object classification, our hand-primed method can be more suitable for real-world object recognition tasks (*e.g.*, teachable object recognizer) that can benefit blind people.

Chapter 6: Hand-Primed Sonification for Teachable Object Recognition on Head-Worn Camera[1]

Supporting blind people to take photos plays an import role in assistive technologies designed for object recognition [2,4,57] and pedestrian detection [171]. Many have explored methods that convert visual information into an accessible modality, such as haptic or sonification, to help blind people take photos. Our early chapter (Chapter 4) also touches on this topic by introducing a sonification feedback mechanism, which detects the object of interest based on its proximity to the user's hand and provides non-visual feedback that informs the blind user of the object location within camera frame. Our first contributions focused on a hand-held camera in smartphones, the most popular platform for assistive technologies.

This chapter revisits our sonification approach to examine how well they generalize on a different form factor, a head-worn camera. Specifically, we explore this in the context of teachable object recognition (keeping the use case scenario as a constant) that is deployed on smartglasses (changing only the form factor). As shown in Chapter 4, blind users take photos both when teaching the machine learning model to recognize a new object and when trying to recognize an object. In our smartglasses implementation of a teachable object recognizer, blind participants do not have to use their hand to aim the camera towards the object of interest. In-

[1]The chapter presents joint work with Professor Hernisa Kacorri, Dr. Jonggi Hong, and graduate students at IAM Lab: Rie Kamikubo, Farnaz Zamiri Zeraati, and Ebrima Jarjue.

stead, their hands would be free to interact with the object of interest. To explore blind user's interaction with our sonification feedback and with this first-of-its-kind wearable teachable object recognizer, we conducted a user study with 13 participants who self-reported as totally blind or legally blind. Our user study was conducted in a semi-deployment manner in that we delivered the study materials including smartglasses and object stimuli to participants' homes and had them evaluate the system in their own environment. Participants first taught the smartglasses to learn three snacks provided by us and three objects of their own. Then, they tested how the system performed in recognizing those objects. Our findings indicate that smartglasses with the sonification feedback can facilitate camera aiming, which is necessary for blind users to interact with their teachable object recognizer. Compared to our previous user study with smartphones, we see that blind participants tend to take photos faster and managed to teach objects as effectively under an uncontrolled environment.

6.1 Motivation

Smartphones and thus hand-held cameras are the status quo for assistive technologies for the blind [2,4,11,17,19,29,32,160,172], though, there are still many challenges related to camera aiming [2,4,11,17,29,31,32,160]. Accessibility researchers investigated methods that can make camera aiming more accessible to blind people so that they can use the camera to recognize their visual surroundings [4, 17, 29, 31, 32]. Indeed, some of these research ideas are quickly translated to real-world applications [10,24,35,36,40,44,47,54,81,173,174]. However, the use of cameras along with computationally intensive tasks can rapidly heat up the phone and drain its battery [175, 176]. Although this issue is not unique to this form factor, it can harm phone

availability that blind users may rely on; many have reported experiencing anxiety when hearing battery warnings in their phones, especially at their work or during commuting [177].

With the promise of hands-free and thus more natural interactions, head-worn cameras are falling on and off the map of assistive technologies for people who are blind or have low vision [23, 65, 66, 178–182]. This can be partially explained by hardware constraints, such as battery life and weight, limiting their commercial availability [183] which can also relate to factors such as cost, social acceptability, privacy, and form factor [16, 171]. Nonetheless, we see a few attempts in commercial assistive products [10, 24, 81, 184], and people with visual impairments remain amicable towards this technology [22], especially when it resembles a pair of glasses [185] similar to the form factor used in our testbed. However, to our knowledge, there is no prior work comparing hand-held camera interaction data from blind people with head-worn camera interaction data from blind people, especially in the context of object recognition. Although wearable cameras do not require wearers to aim a camera, including visual information of interest in the frame can still be challenging for blind people. Prior work with sighted [63] and blind-folded [61] participants indicate that the camera position can lead to different frame characteristics, though such data are not proxy for blind people's interactions nor related to the object recognition scenario. Thus, collecting and analyzing data from blind people, our work takes a step towards understanding their interactions with a head-worn camera in the context of object recognition.

6.2 Methods

6.2.1 System Design

The prototype system consists of two major components: a front-end component and a back-end component. The front-end component is an Android app running on smartglasses and is responsible for capturing images, sending them to the back-end component, and providing sonification based on feedback from the back-end component. The back-end component supports two functionalities—the sonification feedback and teachable object recognition. As for the sonification support, it takes an image as input from the front-end module to an object localization model that estimates an object center area based on the hand information. This center estimation is then sent back to the front-end module that plays sonification for the user accordingly. We provide more detailed information about our prototype system components below.

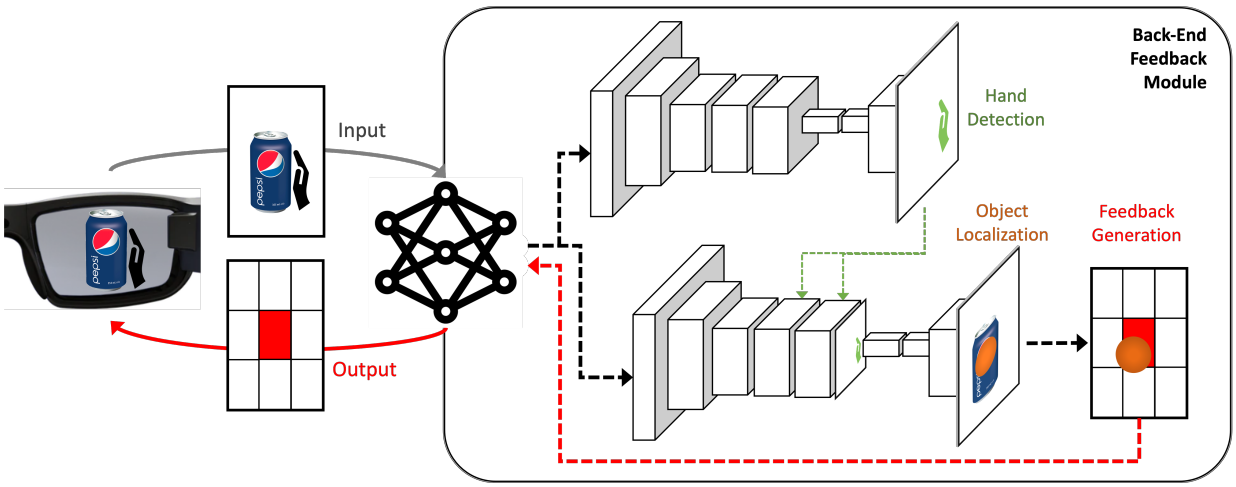


Figure 6.1: System architecture of our sonification feedback designed for smartglasses for the blind. The overall diagram is similar to that in Chapter 4. The main difference is that the smartglasses one is powered by our hand-primed object localization model introduced in Chapter 5 and does not provide the haptic feedback when the object is at the center of camera frame.

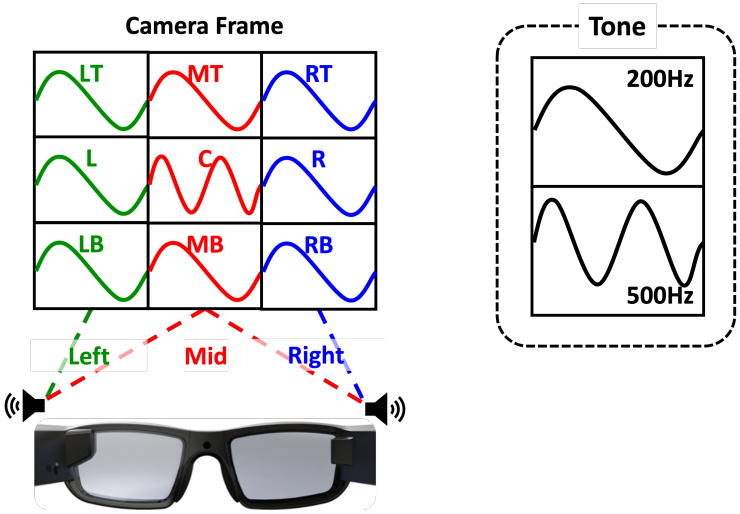


Figure 6.2: Our sonification feedback mechanism designed for smartglasses for blind users.

6.2.1.1 Front-end component: smartglasses app

The main objective of the smartglasses testbed app is to help blind users teach their objects to smartglasses and recognize those objects later when necessary. The smartglasses app was developed with an iterative design method by one sighted researcher and one blind researcher. Based on prior work [4, 171], the researchers first discussed three design goals for the smartglasses app—the practice feature where the user can learn how to use the testbed app, the teaching feature where the user can teach the smartglasses about their objects, and the recognition feature where the user can audit the teachable smartglasses. Starting from the practice feature to the recognition feature, they implemented and improved each of the features iteratively. Throughout this iterative design process, the blind researcher focused on evaluating the testbed app and shared critical feedback from the user perspective, and they ensured accessibility of the testbed app for blind users. The testbed system was developed on Vuzix Blade [186] running on Android but without the TalkBack feature, which is one of the accessibility features available on Android for blind users. To compensate for this accessibility issue, the researchers implemented

a TalkBack-like feature so that the blind user can navigate the menu and interact with the system through audio cues.

The type of input data is an image. To help blind people perceive their camera aiming behaviors, the app is equipped with non-visual feedback, which is sonification. This sonification feedback aims to inform the blind user of an object's location within camera frame, which the user wants to teach to the smartglasses or recognize. The sonification feedback mechanism mostly follows its predecessor introduced in Chapter 4 except for the haptic feedback—this smartglasses version does not provide haptic feedback when the object is estimated at the center of camera frame. Figure 6.2 depicts that the camera frame is divided into nine-equal-size rectangular regions—from the top to bottom, the left three regions are labeled as LT, L, and LB, respectively; the middle three regions are labeled as MT, C, and MB, respectively; and the right three regions are labeled as RT, R, and RB, respectively. Our sonification mechanism uses stereophonic sound to distinguish when the object is on the left, middle, or right of camera frame. In particular, a high-pitch tone sound (*i.e.*, 500Hz sinusoidal wave) is employed to highlight when the object is estimated at the center of camera frame—it plays a low-pitch tone sound (*i.e.*, 200Hz sinusoidal wave) when the object is found to be on the other areas of camera frame. In summary, we apply the following mapping between the region and the feedback modality:

- **LT, L, LB**: left-stereophonic sinusoidal wave (200Hz)
- **MT, MB**: middle-stereophonic sinusoidal wave (200Hz)
- **RT, R, RB**: right-stereophonic sinusoidal wave (200Hz)
- **C**: middle-stereophonic sinusoidal wave (500Hz)

6.2.1.2 Back-end component: remote server

In our architecture, a remote server is necessary to handle heavy-computation tasks, such as hand-guided object localization and teachable object recognition, which can be quickly processed by GPUs. Thus, we set up a server with eight GPUs to facilitate not only object localization but teaching and recognition.

As for object localization, we employed our hand-primed object localization model proposed in Chapter 5. This model consists of two fully convolutional networks (FCNs). One of these FCNs is designed to detect and segment egocentric hand(s) from an image, which is then infused to the conv4 and conv5 layers of the other FCN. The objective of the other FCN is to detect the object of interest to the person whose hands are shown in an image. In other words, this model aims to detect the object of interest (the object's center) according to the user's hand that is also captured in an image. A threshold of 0.5 is used to determine the center area of the object of interest, and the other implementation details are the same as those in Chapter 5.

Object recognition was powered by the Inception [105] V3 model pre-trained on the ImageNet dataset [163]. Its teachable function was implemented by keeping pre-trained features and retraining only its last layer that encodes object labels provided by the user. Object detection models, such as Faster R-CNN [6], tend to result in more accurate object recognition, but they require more annotation data (*i.e.*, object bounding boxes) in addition to object labels. In the context of teachable object recognition for the blind, providing such annotation data during teaching is not accessible, so we chose this image classification model over object detection models. This model is trained with the the following parameters: 32 batch size, 0.01 learning rate, and no data augmentation.

6.2.2 User Study

We designed a multi-session experiment to understand blind people's perception and interaction with AI-infused smartglasses in terms of teachable object recognition [4, 57]. It consists of two sessions—the first session in which we collected participants' background information including demographics; and the second session in which participants use our smartglasses testbed in the context of teachable object recognition. The second session involves three main tasks followed by a survey questionnaire and a semi-structured interview, respectively. Only the second session was recorded for our analysis on participants' qualitative responses.

6.2.2.1 Study procedures

Our user study, which was designed to be conducted remotely, consists of two sessions. In each session, one participant was guided by one experimenter remotely through an audio or video conferencing tool. The following is the description of our study protocol.

Recruitment and Screening: We recruited participants by sending a recruitment email to our participants pool and sharing a recruitment post with a local community for the blind. When recruiting participants, we also asked them to share our study information with their colleagues or friends who they think would be eligible. Potential participants who showed interest in the study were then informed of our consent form and screened based on the criteria of our user study—participants must be at least 18 years old and totally or legally blind. Communication was mainly done through email—sometimes a phone call was employed if participants preferred.

Session 1: Once eligible participants agreed on our consent form, they were invited to the first session. In session 1, we gathered demographics related to age and onset vision impairment

as well as technology attitudes and prior experience with technology such as smartphone, smartglasses, and artificial intelligence. No audio or video recording was done in this session. Instead, the experimenter read the questionnaire took notes on participants' verbal responses. Session 1 took about 30 minutes.

Session 2: At least a day after session 1 was conducted, participants were invited to the second session. Prior to session 2, another experimenter delivered the study equipment including smartglasses with a portable charger, a portable WiFi hotspot device, and five grocery products, and, if they asked, helped participants set up the study environment in their preferred place.

In session 2, participants went through four different phases (practice, teach 1, test 1, and teach 2 & test 2). Using the Practice mode available on GlAccess and two objects (a coffee pod and a soda can) we provided, they first learned how to record an object label and take photos on smartglasses. In the next two phases (teach 1 and test 1), participants were asked to teach and test GlAccess, respectively, using the other three objects we delivered to them—a bag of Cheetos, a bag of Fritos, and a bag of Lays. While the Teach mode specifically instructed them to take 30 photos for each object, participants freely decided the number of photos to test their personalized GlAccess. After these two phases, we asked participants to choose their own three objects to teach and test GlAccess. No constraints were indicated for their choice of objects, and no specific order between teaching and testing was instructed—that is, participants can choose to teach GlAccess about all their three objects and test it with those objects or they can teach and test GlAccess using each object separately.

Table 6.1: Participants' demographics and prior knowledge with machine learning. Light perception is indicated with an asterisk. Some demographic information for P2 and P9 are not presented here, which is marked with dash (-), due to their request.

ID	*Age*	*Gender*	*Level of vision*	*Age of onset*	*Machine learning*
P1	59	Male	Legally blind	1	Somewhat familiar
P2	-	-	Totally blind*	Birth	Somewhat familiar
P3	72	Female	Totally blind*	Birth	Extremely familiar
P4	67	Male	Totally blind*	52	Slightly familiar
P5	48	Male	Totally blind	29	Slightly familiar
P6	32	Female	Legally blind	21	Somewhat familiar
P7	53	Female	Totally blind	37	Somewhat familiar
P8	67	Female	Totally blind*	Birth	Slightly familiar
P9	-	-	Totally blind	Birth	Somewhat familiar
P10	37	Female	Totally blind	3	Somewhat familiar
P11	52	Female	Legally blind	47	Somewhat familiar
P12	71	Male	Legally blind	Birth	Somewhat familiar
P13	67	Male	Totally blind*	50	Extremely familiar

6.2.2.2 Participants

As Table 6.1 presents, 13 participants who are either totally blind or legally blind participated in our user study (IRB #1822836-3). Four participants self-identified as legally blind, and the other eight participants as totally blind—among the totally blind participants, five reported having light perception. Their ages ranged from 32–72 (μ: 56.8, σ: 13.7)—P2 and P9's age and gender are not reported according to their request. None of them had used smartglasses in their daily lives. All of them had some knowledge about machine learning—two reported having a solid understanding of that concept.

6.3 Results

We employ quantitative methods as well as qualitative methods to analyze our study data. The quantitative methods involve measuring accuracy in object recognition and word error rate

(WER) in sonification feedback. Several qualitative methods, such as clustering analysis and image coding, are adopted to observe some patterns from the data and understand the data in a quantitative way. The image coding is performed by one sighted researcher who follows the image coding scheme developed by Hong *et al.* [133].

6.3.1 Object Recognition Performance

We first investigate the feasibility of teachable object recognition on smartglasses by assessing accuracy of models taught by participants. Since they taught their smartglasses to recognize two different sets of objects (the three snacks provided by us and three other objects chosen by themselves), we evaluate their object recognizers on these two different sets, respectively. In this evaluation, the set of three snacks from us is categorized as *snacks* while the other of their three own objects as *theirs* as shown in Figure 6.3. More specifically, their models were trained on all images of the snacks and their own objects (*i.e.*, 30 images of each object) and tested on images of the snacks and theirs respectively—the number of testing images varies by participant and object as participants were free to decide it on their own.

Figure 6.3 shows the average accuracy of participants' object recognizers on two different

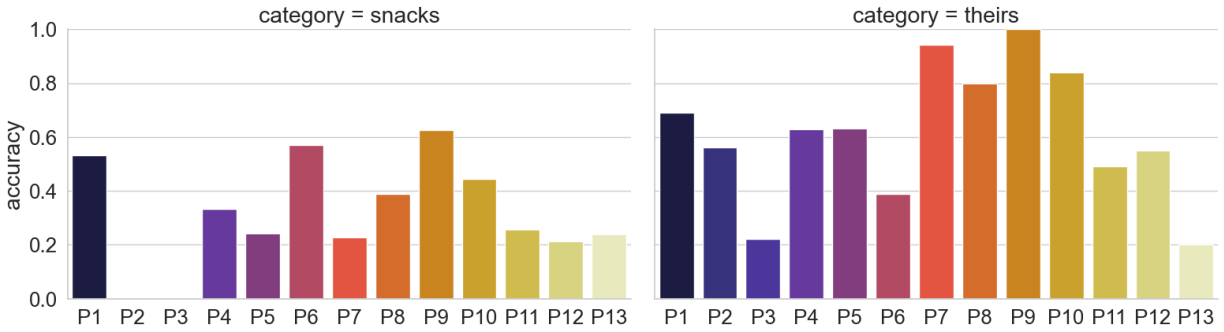


Figure 6.3: Average model accuracy per participant.

sets of objects. Overall, recognition of their own objects tend to be higher than that of the three snacks that were our object stimuli. This is not surprising. The three snacks were specifically engineered to serve as difficult case of a fine-grained recognition, with objects being similar in size, shape, and color as well as being deformable with reflective surfaces. Some promising results are worth to be noted. Object recognizers taught by some blind participants (*e.g.*, P7 and P9) achieve higher accuracy than the other participants. In their testing images, we see that their own objects are clearly visible and different in shape and color from one to another, which may have helped their object recognizers learn visual features that distinguish them. Given that all participants had no prior experience with smartglasses (camera aiming in particular) and taught smartglasses in their houses (*i.e.*, real-world environments), it is promising to see that smartglasses taught by seven participants (P1, P4, P5, P7, P8, P9, and P10) achieve more than 60% accuracy. In particular, four among them (P7, P8, P9, and P10) managed to teach object recognizers that can achieve similar accuracy or higher than 80%.

Moreover, when choosing their own three objects (*i.e.*, "theirs" category in Figure 6.3), we observe that participants opt for objects that are different in shape, making it easier to distinguish from one another. This difference explains some of the performance gap between these sets of objects (*i.e.*, snacks vs. theirs). Yet, there are several factors that could explain the low model performance here beyond how distinguishable the objects are. For example, it could be that participants' training examples were not sufficient enough for the model to learn to distinguish those objects. Or it could be that the models are well trained but the photos that blind participants took and used to test their object recognizers are not capturing the objects well. To decouple these factors we perform a detailed analysis based on visual inspection of participants photos. In this visual inspection, one sighted experimenter reviewed the training data set while two other sighted

experimenters annotated the testing data set to check if the testing data contains enough visual information for a sighted person to perform object recognition without any additional context. The testing dataset was randomized before being given to the two sighted experimenters. The experimenters were provided with a set of object labels. Thus, their annotation is based on the closed-world assumption, similar to the classification model for the object recognition task. That is, during the annotation process, the experimenters looked at each image and chose an object label from a given list of labels assigned by the participant who took that image. In this annotation, the experimenters were instructed to try to not use any context (*e.g.*, previous images) to "infer" an object label. Instead, they were asked to label an object when they are sure about the object label—that is, when an image contains enough visual information that any sighted person without context will be able to recognize the object.

Our visual inspection reveals three cases—(i) testing data did not include necessary visual information; (ii) testing data did include necessary visual information, but training data did not; and (iii) both training data and testing data did include necessary visual information, but the model was unable to recognize the object. In this inspection, a data set was considered to contain necessary visual information if all the experimenters managed to tell the right object label in more than 60% of the dataset (*i.e.*, more than 0.6 accuracy)

One representative example for the first case is P3's testing data for the snack objects. P3's testing images do not capture the target objects at all or just very partially with annotations from sighted experimenters, indicating that one would not be able to tell what this object is just from the photo. More specifically, P3 took only one or two images for testing, but the target object is not included or mostly occluded by their hand. As a result, the computer vision model was also unable to provide correct recognition results given those images. Another example is P11, who

during the testing took photos of objects from far away. Thus, those objects were captured as too small. Detecting and recognizing a small object is another well-known challenging task in computer vision [187].

Regarding the second case, we found that P3 and P13 were not able to provide good training images. Their object recognizers achieved less than 30% accuracy although their testing images included necessary visual information for recognition. From their training, we find a common pattern that their target objects are either occluded, partially captured, or not included at all. More specifically, in all 90 training images captured by P3 (30 images for each of P3's three objects), the target objects are partially include or occluded by P3's hair. Among 90 training images from P13, 87 images do not include P13's target objects, and only three images include part of the objects. We suspect that excluding the objects' distinct visual features may have hampered the learning process and thus affected the recognition accuracy.

P2's testing images for the snacks objects fall under the third case. P2's testing data included the target objects, but its personalized object recognizer was confused about the snack objects with one of P2's own objects, which is a different variety of the Cheetos snack. It led the model to failing to recognize all the snack objects. It is noteworthy that introducing a similar object could lead to failure of recognizing other similar objects in the context of teachable object recognition.

6.3.2 Camera Aiming Behaviors

To learn blind participants' camera aiming behaviors with smartglasses, we analyze their images, particularly captured during the teaching phase. Sonification feedback data and im-

ages taken by blind participants are formed into a stream of interactions for a participant–object pair—three feedback data received right before each photography are used to preserve the temporal information in their camera interactions. Since 13 participants taught their smartglasses to recognize 6 objects, 78 interactions streams are formed and used in this analysis. As in Chapter 4, a stream consists of 30 four-element-tuples (f_3, f_2, f_1, Take), where f_i could be:

- *None*: No feedback generated (no object location was estimated by our localization model).
- *Center*: The high-tone and haptic feedback generated (an object was estimated to appear at the center, region **C** in Figure 6.2).
- *Left*: The left-stereophonic tone feedback generated (an object was estimated to appear on region **LT**, **L**, or **LB**).
- *Right*: The right-stereophonic tone feedback generated (an object was estimated to appear on region **RT**, **R**, or **RB**).
- *Middle*: The mid-stereophonic tone feedback generated (an object was estimated to appear on region **MT** or **MB**).

The Louvain method for community detection algorithm [137] is used to find some common patterns from these interaction streams. Refer to Chapter 4 (Section 4.3.3) for the detailed implementation of this clustering analysis. The clustering algorithm identifies three clusters as depicted in Figure 6.4.

In this analysis, we focus on understanding blind people's camera aiming behaviors with smartglasses compared to theirs with smartphones, which is shared in Chapter 4.3.3. The descriptive features of each cluster in Figure 6.4 are mostly similar to those in Figure 4.12. This

clustering result shows three distinct behaviors (C1, C2, and C3). The C1 cluster, to which 30 participant–object pairs belong, shows a case where participants received no sonification feedback. 29 participant–object pairs fall under the C2 cluster representing participants' tendency of positing their objects at the center of camera frame. The third cluster (C3) represents 19 participant–object pairs where participants received any other sonification feedback before taking photos. One particular pattern that is not observed in the interactions with smartglasses is that participants did not tend to hear the "left" sonification feedback—that is, the object of interest was not located on the left side of camera frame. Two factors might have potentially led to this pattern. The first one is related to this specific camera form factor where camera is located on the left side. The second is related to blind participants' tendency of locating the object of interest in front of their face—that is, their smartglasses. Some participants moved objects around in front of their smartglasses to receive sonification feedback and know its location within the camera frame. The lack of the "left" sonification pattern in their interactions indicates that participants took photos whenever they hear some sonification feedback, which tends to be either "center," "middle," or "right."

We then project participant–object pairs onto the participant–object space with their clus-

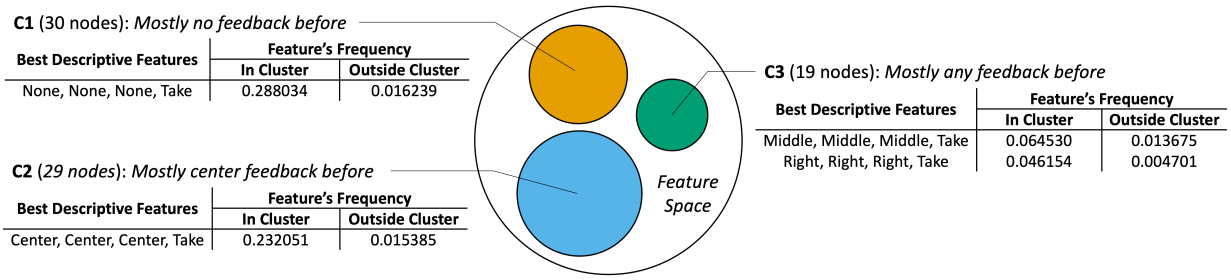


Figure 6.4: By representing the stream of interactions in a (participant, object) pair as a unique node, we find three distinct clusters and their descriptive features that are most responsible for this cluster formation.

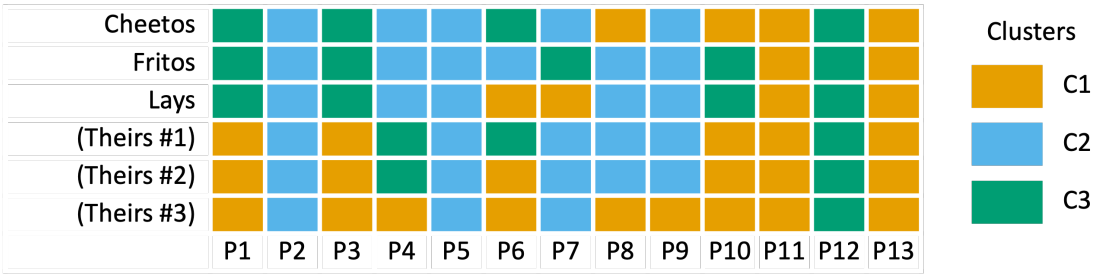
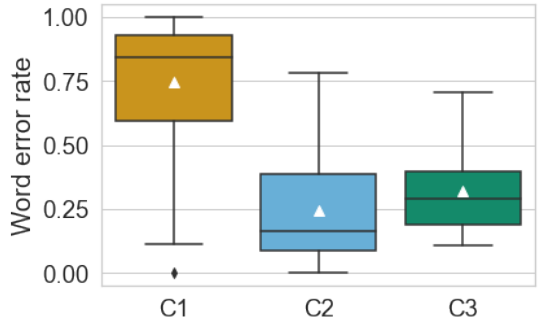


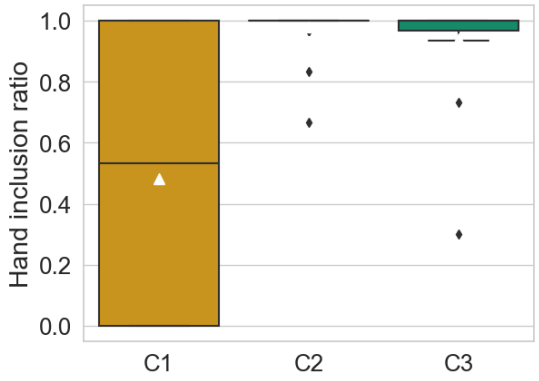
Figure 6.5: Cluster distribution across participant-object pairs.

tering results as shown in Figure 6.5. This cluster distribution allows us to observe participants' interactions with the sonification feedback per object. Overall, participants tended to receive the same type of sonification feedback across their objects. P11 and p13 tended to not receive any feedback before taking photos. While our visual inspection already reveals that P13 did not include objects within camera frame during the teaching and thus did not hear any sonification feedback, we observe different issues from P11's images. P11 captured the three snacks under insufficient lighting. P11's own objects were under a better lighting condition but occluded by the hair or captured too small—it seems that P11 sometimes put the objects on the floor and took photos while standing. These issues found from their teaching images are still observed from those streamed for sonification feedback. Similar issues were observed from other participant–object pairs belonging to the C1 cluster. For example, when P1 took photos of P1's own objects for teaching, P1 tended not to hold those objects (no hand inclusion is observed from 71 photos out of 90 teaching photos). The sonification feedback mechanism is designed to estimate the object location according to hand information (*i.e.*, hand pose and location), so data without the hand information seems to be a challenging case to the feedback mechanism.

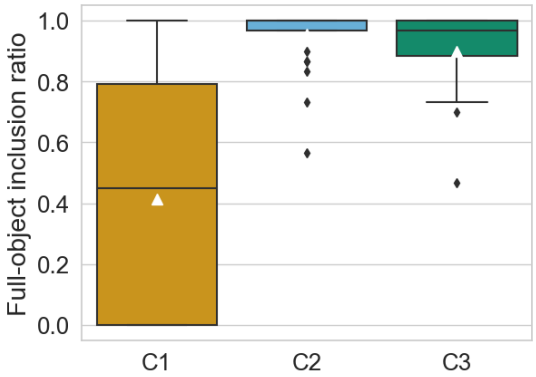
From participants' interactions with the sonification feedback on smartphones, we observe two distinct types of interactions: *participant-specific* and *object-specific* (reported in Sec-



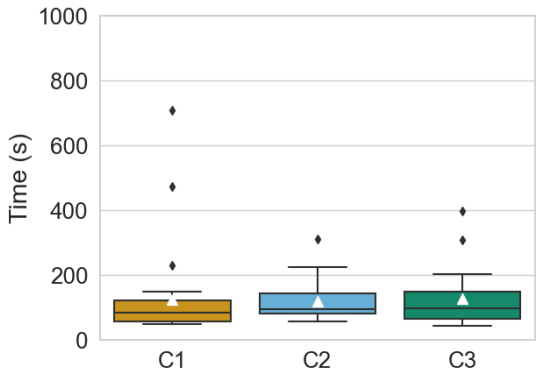
(a) Feedback word error rate (WER) across pairs.



(b) The ratio of photos including hands.



(c) The ratio of photos including the object fully.



(d) Time spent in teaching.

Figure 6.6: Comparing clusters across different interaction characteristics during teaching.

tion 4.3.3). Unlike our previous study with smartphones in which participants used 15 objects, only the three snack objects were commonly used by participants in our study with smartglasses. Hence, the study setup itself is not feasible to observe object-specific interactions (*i.e.*, similar patterns in interactions with a certain object across participants). However, participant-specific interactions are still observed from our participants' smartglasses camera use. Figure 6.5 shows that participants tended to interact with the same type of sonification feedback (*i.e.*, the same camera aiming behavior) when teaching all objects (P2, P5, P9, P11, P12, P13) or the three snacks at least (P1, P3, P4, P10).

To understand the clusters further, we analyze these clusters with respect to four differ-

ent characteristics: feedback error, the ratio of photos including hands, the ratio of photos including the whole object, and time spent for teaching. First of all, participants using smartglasses tended to teach objects faster than those who taught objects to smartphones throughout all different clusters—in Figure 6.6(d): C1 (mean=123.57, std=136.80), C2 (mean=118.72, std=57.17), C3 (mean=126.11, std=92.32) vs. in Figure 4.13(d): C1 (mean=227.94, std=140.32), C2 (mean=220.28, std=194.55), C3 (mean=129.90, std=65.63). Given that participants from both of the studies had no or little photo-taking experience with their associated camera form factor, it seems that photography on smartglasses may be less time-consuming than that using smartphones. It may indicate that wearing camera may reduce the burden of camera aiming, but more research is needed to understand this further. Figure 6.6(b) presents that participants received the center feedback (C2) or any other feedback (C3) when their hand is in proximity to the object, which is the case that our sonification feedback mechanism works the best. Similarly, participants interacted with the sonification feedback when they captured the whole object, as shown in Figure 6.6(c). In particular, when they heard the center feedback, they tended to capture the whole object. The high variance observed from the C1 cluster in Figure 6.6(b) and (c) shows that participants did not hear any feedback since their hand did not appear near their target objects or did not include their objects appropriately. However, at the same time, it also shows that sonification feedback was not provided even though they did hold their objects with their hand or include their whole objects within camera frame. The C1's word error rate (mean=0.75, std=0.25) presented in Figure 6.6(a) indicates it in that sonification feedback was not provided when participants indeed included their objects within camera frame, which was, on average, 75% of the times. The high error rate of the C1 cluster suggests that more research is needed to improve the recall performance of machine learning models.

6.4 Discussion

Smartglasses show great promise in helping blind users to access their everyday objects, but more work is necessary to design such assistive systems properly. In the following, implications and limitations of our findings are discussed to facilitate future research.

6.4.1 Implications

We observe that some blind participants managed to achieve competitive or higher accuracy in object recognition than 80%. However, this performance achievement could be susceptible to objects they chose and visual similarities or differences among those objects. For example, P7, who achieved more than 90% accuracy in recognizing her own three objects, ended up having less than 30% accuracy in recognizing the three snacks provided by us, which are engineered to serve as a worse case scenario as they are visually similar deformable objects with reflective surfaces. Multiple factors may have played a role in creating this discrepancy and thus need to be studied further to realize object recognition for blind users. One factor might be related to the fact that our evaluation uses an image classification model, not an object detection model that is more desirable for teachable object recognition. As presented in Chapter 5, providing more explicit information about the object of interest (*i.e.*, bounding box) can help a computer vision model to learn distinguishable features of the object and thus better recognize it. However, drawing a bounding box is difficult and time-consuming to be done in a real-time manner [188]. Even more, this task itself requires a visual perceptual skill as it involves "drawing" a bounding box around the object, which would not be accessible to blind users. Employing object detection models instead of image classification models is promising in that it can help to improve the performance

of teachable object recognition, but future research needs to investigate how to help blind users provide such additional information when teaching their teachable object recognizer.

Smartglasses unlock new interactions, which are not possible with smartphones. Hands-free interactions can be an eminent example. Our findings indicate that this head-worn camera form factor can help blind users capture their target object with less effort (Figure 6.6) than that with smartphones (Figure 4.13). However, there are still cases in which some blind users captured only part of the object or did not include the object at all within camera frame, although they thought they were capturing the object. It indicates that more tools are needed to help blind users perceive their camera aiming behaviors. Providing additional descriptors about visual information captured within camera frame may help blind users create their mental mapping between their camera aiming and the camera frame. For example, recent work shows that blind users find it helpful to have visual information regarding whether the target object is cropped or their hand is included within the camera frame [189].

6.4.2 Limitations

We delivered the study materials to blind participants and conducted our user study remotely so that blind participants can freely interact with teachable object recognition on smartglasses in their environments. This deployment setup helped us to alleviate concerns related to the recent COVID-19 pandemic [190] and to evaluate the teachable interface in a more real-world environment than an in-lab setting. However, it did not allow us to observe blind users' natural interactions with such teachable interfaces. In particular, the last phase of Session 2 was designed to mimic a deployment setup that would help us observe blind users' real-world interactions. For

example, we had expected that blind participants would teach and test their smartglasses with their own objects around the objects' original locations, but many participants brought their objects back to the place where they communicated with the remote experimenter. We suspect that how blind users interact with teachable smartglasses for object recognition might be different from what we observed from our remote user study. Furthermore, blind participants were asked to choose only three objects of their own in this phase. Some participants chose three objects that are visually similar to one another while some others picked up three objects that do not share common visual characteristics. It indeed prevents us from learning the distribution of objects that blind users want such a teachable interface to help them recognize. A long deployment study is necessary to obtain real-world data, showing how blind users use such teachable smartglasses and what kind of objects they need an assistive system to help recognize in their daily lives.

6.5 Summary

In this chapter, we developed teachable smartglasses for blind people and evaluated this assistive system for object recognition in real-world environments (*i.e.*, participants' own environments, not in a lab environment). The sonification feedback module, which was initially designed to help blind people aim their smartphone camera towards the object of interest, was also employed to help blind users create their mental mapping between their interactions with the object and the object's location within camera frame. Our findings suggest that smartglasses have its potential in facilitating camera aiming towards the region of their interest. Comparing to our previous observations from the user study using smartphones in the same context, we see that blind participants tended to perform the teaching process faster on smartglasses and use their

hand more often to capture the object of interest within camera frame. Indeed, it suggests that, with smartglasses, blind users can easily use their hand to not only include the object within camera frame but also provide a cue to the object of interest, which can help computational models to detect it among other objects captured together. We envision that smartglasses that enable users to move their hands freely can unlock various user interactions, which are not possible with the currently universal camera form factor (*i.e.*, smartphones).

Chapter 7: Blind People's Interactions with Pedestrian Detection System on Head-Worn Camera[1]

Access to the spatial behavior of passersby can be critical to blind individuals. Passersby proxemic signals, such as change in distance, stance, hip and shoulder orientation, head pose, and eye gaze, can indicate their interest in "initiating, accepting, maintaining, terminating, or altogether avoiding social interactions" [192,193]. More so, awareness of others' spatial behavior is essential for preserving one's personal space. Hayduk and Mainprize [194] demonstrate that the personal space of blind individuals does not differ from that of sighted individuals in size, shape, or permeability. However, their personal space is often violated as sighted passersby, perhaps in an attempt to help, touch them, or grab their mobility aids without consent [195, 196]. Last, health guidelines as with the recent COVID-19 pandemic [190] require one to practice social distancing by maintaining a distance from others of at least 3 feet (1 meter)[2] or 6 feet (2 meters)[3], presenting unique challenges and risks for the blind community [199, 200]; mainly, due to the fact that spatial behavior of passersby and signage or markers designed to help maintain social distancing are predominantly accessible through sight and thus are, in most cases, inaccessible to blind individuals. Other senses such as hearing and smell could be utilized perhaps to estimate

[1]This chapter presents joint work with Professors Hernisa Kacorri and Chieko Asakawa, Dr. Daisuke Sato, and Saki Asakawa published in ACM ASSETS 2021 [191].

[2]Minimum distance suggested by the World Health Organization (WHO) [197]

[3]Minimum distance suggested by the Center for Disease Control and Prevention (CDC) in the United States [198]

passersby distance and orientation but require close proximity or quiet spaces; in noisy environments, blind people's perception of surroundings out of the range of touch is limited [201]. More so, mobility aids may work adversely in some cases (*e.g.*, guide dogs are not trained to maintain social distancing [202]).

Why computer vision and wearable cameras. Assistive technologies for the blind that employ wearable cameras and leverage advances in computer vision, such as pedestrian detection [61, 171, 203, 204], could provide access to spatial behavior of passersby to help blind users increase their autonomy in practicing social distancing or initiating social interaction; the latter motivated our work since it was done right before COVID-19. 'Speak up' and 'Embrace technology' are two of the tips that the LightHouse Guild provided to people who are blind or have low vision for safely practicing social distancing during COVID-19 [199], mentioning technologies such as Aira [24] and BeMyEyes [40]. However, these technologies rely on sighted people for visual assistance. Thus, beyond challenges around cost and crowd availability, they can pose privacy risks [85, 205, 206]. Prior work, considering privacy concerns for parties that may get recorded, has shown that bystanders tend to be amicable toward assistive uses of wearable cameras [16, 82], especially when data are not sent to servers or stored somewhere [171]. However, little work explores how blind people capture their scenes with wearable cameras and how their data could work with computer vision models. As a result, there are few design guidelines for assistive wearable cameras for blind people.

To better understand the opportunities and challenges of employing such technologies for accessing passersby proxemic signals, we collect and analyze video frames, log data, and open-ended responses from an in-person study with blind (N=10) and sighted (N=40) participants[4]. As

[4]The data were collected right before COVID-19 with a use case scenario of blind individuals initiating social

shown in Figure 7.1, blind participants ask passersby for information while wearing smartglasses with our testbed prototype for real-time pedestrian detection. We explore what visual information about the passerby is captured with the head-worn camera; how well pedestrian detection algorithms can extract proxemic signals in camera streams from blind people; how well the real-time estimates can support blind users at initiating interactions; and what limitations and opportunities head-worn cameras can have in terms of accessing proxemic signals of passersby.

In this chapter, we present our exploratory study showing that there are still many limitations to overcome for such assistive technology to be effective. When the pedestrians faces are included, the results are promising. Yet, idiosyncratic movements of blind participants and the limited field of view in current wearable cameras call for new approaches in estimating passersby proximity. Moreover, analyzing images from smartglasses worn by blind participants, we observe variations in their scanning behaviors (*i.e.*, head movement) leading to capturing different body parts of passersby, and sometimes excluding essential visual information (*i.e.*, face), especially when they are near passersby. Blind participants' qualitative feedback indicates that they found it easy to access proxemics of passersby via smartglasses as this form factor did not require camera aiming manipulation. They generally appreciated the potential and importance of such an wearable camera system, but commented on how experience degraded with errors. They mentioned trying to adapt to these errors by aggregating multiple estimates, instead of relying on a single one. At the same time, they shared their concern that there was no guaranteed way for them to check errors in some visual estimates. These findings help us discuss the implications,

interactions with sighted people in public spaces. Out team, co-led by a blind researcher, was annotating the data when the pandemic reached our shores. Thus, we felt it was imperative to expand our annotation and analysis to include distance thresholds recommended for social distancing. After all, both scenarios share common characteristics such as need for proximity estimation and initiating interactions to maintain distancing, *i.e.* recommended 'Speak up' strategy by the LightHouse Guild [199].

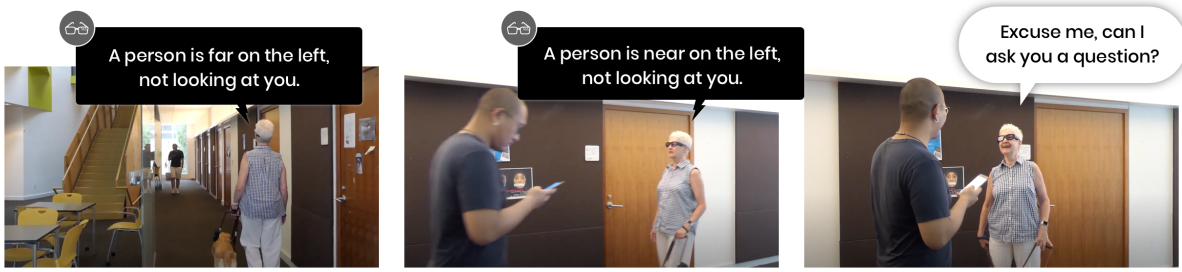


Figure 7.1: A pedestrian detection scenario demonstrating the data collection in our study. A blind person, wearing smartglasses with our working prototype, and a sighted person walk toward each other in a corridor. The smartglasses detect the sighted passerby, estimate his proximity, and share his relative location and head pose.

opportunities, and challenges of an assistive wearable camera for blind people.

7.1 Motivation

Our work is informed by prior work in assistive technology for non-visual access of passersby spatial behaviors and existing approaches for estimating proximity beyond the domain of accessibility. For more context, we also share prior observations related to blind individuals' interactions with wearable and handheld cameras.

Prior work on spatial awareness for blind people with wearable cameras and computer vision has mainly focused on tasks related to navigation [21, 207–210] and object or obstacle detection [207, 209, 211–213]. We find only a few attempts exploring the spatial behavior of others whom a blind individual may pass by [214,215] or interact with [60,61,216]. For example, McDaniel *et al.* [214] proposed a haptic belt to notify blind users of the location of a nearby person, whose face is detected from the wearable camera. While closely related to our study, blind participants were not included in the evaluation of this system. Thus, there are no data on how well this technology would work for the intended users. Kayukawa *et al.* [215], on the

other hand, evaluate their pedestrian detection system with blind participants. However, the form factor of their system, a suitcase equipped with a camera and other sensors, is different. Thus, it is difficult to extrapolate their findings for other wearable cameras such as smartglasses that may be more sensitive to blind people's head and body movements. Designing wearable cameras for helping blind people in social interactions is a long-standing topic [216–219]. Focusing on this topic, our work is also inspired by Stearns and Thieme [61]'s preliminary insights on the effects of camera positioning, field-of-view, and distortion for detecting people in a dynamic scene. However, they focused on interactions during a meeting (similar to [60]) and analyzed video frames captured by a single blindfolded user, which is not a proxy for blind people [220–223]. In contrast to our work, the context of a meeting does not allow for extensive movements and distance changes—typical tasks when proxemics of passersby are needed. In this chapter, we explore how blind people's movements and distances can lead to visual characteristics of their camera frames.

7.2 Methods

7.2.1 System Design

We build a testbed called *GlAccess* to explore what visual information about passersby is captured by blind individuals with a head-worn camera and how well estimates of passersby's proxemics can support initiating interactions. As shown in Figure 7.2, GlAccess consists of Vuzix Blade smartglasses [186], a Bluetooth earphone, and a computer vision system running on a server. With a sampling rate of one image per second, a photo is sent to the server to detect the presence of a passerby and extract proxemic signals, which are then communicated to the user

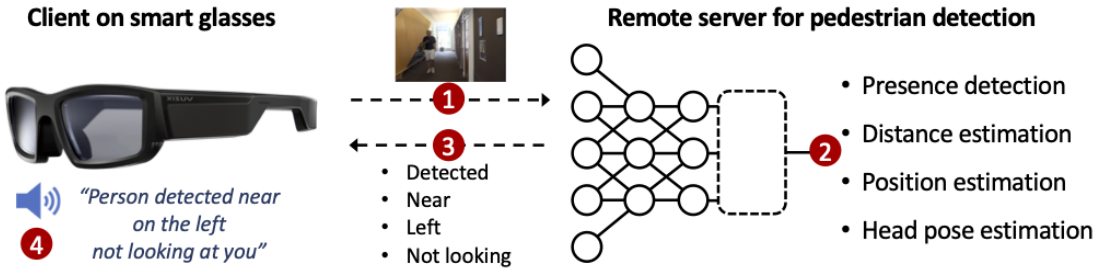


Figure 7.2: System diagram describing the procedure of pedestrian detection: (1) The client app on smartglasses captures an image and sends it to the remote server; (2) Once the remote server detects a passerby, it estimates the distance, position, and head pose of the passerby; (3) The output is then sent back to the smartglasses; (4) The smartglasses deliver the output to the blind user via audio.

through text-to-speech (TTS). To mitigate cognitive overload, estimates that remain the same as the last frame are not communicated. Vuzix Blade smartglasses have an 80-degree diagonal angle of camera view and run on Android OS. Our server has eight NVIDIA 2080 TI GPUs running on Ubuntu 16.04. We build a client app on the glasses that is connected to the server through RESTful APIs implemented with Flask [224] and uses the IBM Watson TTS to convert the proxemic signals to audio (*e.g.*, *"a person is near on the left, not looking at you."* as shown in Figure 7.1). Our computer vision system estimates the following related to proxemics:

Presence: A passerby presence is estimated through face detection. We employ a Multi-task Cascaded Convolutional Networks (MTCNNs) [225] pre-trained on 393,703 faces from the WIDER FACE dataset [226], which is a face detection benchmark with a high degree of variability in scale, pose, occlusion, facial expression, makeup, and illumination. The model provides a bounding box where the face is detected in the frame.

Distance: For distance estimation, we adopt a linear approach that converts the height of the bounding box from the MTCNNs face detection model to the distance. Using the Vuzix Blade smartglasses, we collect a total of 240 images with six data contributors from our lab.

Each contributor walks down a corridor that is approximately 66 feet (20 meters); this is the same corridor where the study takes place. We take two photos every one meter for a total of 40 photos. We train a linear interpolation function with the height of faces extracted from these photos and the ground truth distances between the contributors and the photo-taker with the smartglasses. This function estimates the distance as a numeric value, which is communicated either more verbosely (*e.g.*, *'16 feet away'*) or less so by stating only *'near'* or *'far'* based on a 16 feet (5 meters) threshold[5]. Participants in our study are exposed to both (*i.e.*, experiencing each for half of their interactions).

Position: We estimate the relative position of a passerby based on the center location of the bounding box of their face detected by the MTCNNs model. A camera frame is divided in three equal-width regions, which are communicated to the user as *'left'*, *'middle'*, and *'right'*.

Head Pose: We use head pose detection to estimate whether a passerby is looking at the user. More specifically, we build a binary classification layer on top of the MTNNs face detection model and fine-tune it with the 240 images from our internal dataset collected by the six contributors and 432 images from a publicly available head pose dataset [227]. The model estimates whether a passerby is looking at the user or not.

We also simulate future teachable technologies [3,228] that can be personalized by the user providing faces of family and friends. In those cases, the name of the person can be communicated to the user. Sighted members in our group volunteered to serve this role in this study:

Name: We pre-train GlAccess with a total of 144 labeled photos from five sighted lab members so that the system can recognize them in the study setting to simulate a personalized

[5]The 6 feet (2 meters) threshold recommended from CDC for social distancing was not communicated to blind participants in the study as it preceded the COVID-19 pandemic. However, this threshold was added later in the data annotation phase and analysis as it may prove informative for future researchers working towards accessible social distancing systems.

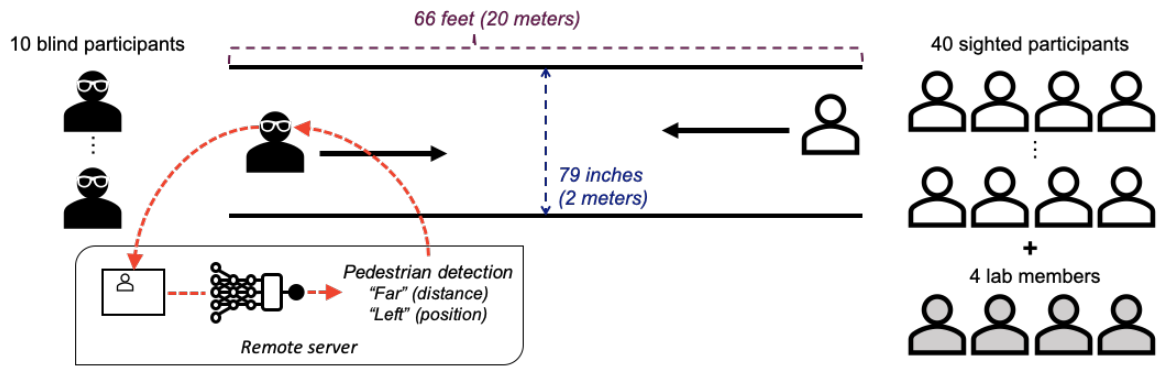


Figure 7.3: Diagram of study procedure. A blind participant walked eight times in a corridor that is 66 feet (20 meters) long and 79 inches (2 meters) wide. Each time a different sighted person walked towards them. In four of the eight times, those individuals were sighted study participants. In the other four, they were sighted members in our research lab. In addition to the proxemic signals, GlAccess can recognize the lab members, simulating future teachable applications [3] where users can personalize the model for familiar faces.

experience. This face recognition is implemented with a SVM classifier that takes face embedding data from the MTCNNs model.

7.2.2 User Study

To explore the potential and limitations of head-worn cameras for accessing passersby proxemics, we conducted an in-person study with blind and sighted participants (under IRB #$STUDY$2019_00000294). We employed a scenario, illustrated in Figure 7.3, where blind participants wearing GlAccess were asked to initiate interactions with sighted passersby in an indoor public space. Given the task of asking someone for the number of a nearby office, they walked down a corridor where sighted participants were coming in the opposite direction.

Table 7.1: Demographic information of our participants. Participants with light perception are marked with an asterisk in *Vision level*.

PID	Gender	Age	Vision level	Onset age	Experience with wearable cameras
P1	Female	71	Totally blind	8	Aira (1 year 6 months; a few times a month)
P2	Male	66	Totally blind	Birth	Audio-glasses (4 months; twice a month)
P3	Male	59	Totally blind	Birth	None
P4	Female	66	Totally blind*	49	None
P5	Female	65	Totally blind	Birth	None
P6	Female	73	Totally blind	Birth	None
P7	Female	49	Totally blind	14	None
P8	Male	60	Legally blind*	56	Aira (1 month; once a month)
P9	Male	56	Totally blind	26	No
P10	Male	71	Totally blind*	41	Aira (3 months; a few times a week)

7.2.2.1 Participants

We recruited ten blind participants; nine were totally blind, and one was legally blind. On average, blind participants were 63.6 years old (SD = 7.6). Five self-identified as female and five as male. Three participants (P4, P8, P10) mentioned having light perception, and four participants reported having experience with wearable cameras such as Aira [24] with frequencies shown in Table 7.1. Responses to technology attitude questions[6] indicate that they are positive about technology in general ($mean$ = 2.02, SD = 0.42) and wearable technology such as smartglasses ($mean$ = 1.89, SD = 0.90).

We also recruited 40 sighted participants to serve as passersby in the study—four sighted participants, one at a time, passed by a blind participant. On average, sighted participants were 25.3 years old (SD = 4.8). Twenty-four self-identified as male and fifteen as female, and one chose not to disclose their gender. Participants' responses to technology attitude questions[6] indicate that they were slightly positive about wearable technology ($mean$ = 1.08, SD = 0.95).

[6]Questionnaire based on [132] is available at `https://iamlabumd.github.io/assets2021_lee`: -3 indicates the most negative, 0 neutral, and 3 the most positive.

7.2.2.2 Procedure

We shared the consent form with blind participants via email to provide ample time to read it, ask questions, and consent prior to the study. Upon arrival, the experimenter asked the participants to complete short questionnaires on demographics, prior experience with wearable cameras, and attitude towards technology. The experimenter read the questions; participants answered them verbally. Blind participants first heard about the system and then practiced walking with it in the study environment and detecting the experimenter as he approached in the opposite direction. During the main data collection session, blind participants were told to walk in the corridor and ask a person, if detected, about the office number nearby. On the other hand, sighted participants, one at a time, were told to simply walk in a corridor with no information about the presence of blind participants in the study. As illustrated in Figure 7.3, a blind participant walked in the corridor eight times—four times with four sighted participants, respectively, and four times with four sighted members of our research lab, respectively. Our study was designed not to extend beyond 2 hours.

All sighted participants were recruited on site and thus strangers to blind participants; blind participants did not meet any sighted participants before the study. A stationary camera was used to record the dyadic interactions in addition to the head-worn camera and the server's logs. Both blind and sighted participants consented to this data collection prior to the study—sighted participants were told that there will be camera recording in the study, which may include their face, but these images will not be publicly available without being anonymized. The study ended with a semi-structured interview for blind participants with questions[6] created for eliciting their experience and suggestions towards the form factor, proxemics feedback delivery, error interaction,

Table 7.2: Attributes in our annotation.

	Presence	**Position**	**Distance**	**Head pose**	**Interaction**
Annotation	Which body part of the passerby is included?	Where is the passerby located?	How far is the passerby from the user?	Is the passerby looking at user?	Who is speaking?
Value	head/torso/arms/legs	left/middle/right	<6ft./near/far	yes/no	blind/sighted

and factors impacting future uses of such head-worn cameras. Their responses except for their qualitative feedback are analyzed in the next chapter.

7.2.2.3 Data annotation

We first annotated all camera frames captured by the smartglasses worn by blind participants. As shown in Table 7.2, annotation attributes include the presence of a passerby in the video frame, their relative position and distance from the blind participant, whether their head pose indicates that they were looking towards the participant, and whether there was an interaction. Annotations for the binary attributes, *presence*, *head pose*, and *interaction*, were provided by one researcher as there was no ambiguity. For *presence*, the annotator marked whether the *head*, *torso*, *arm(s)*, and *leg(s)* of the sighted passerby were captured in the camera frames. Regarding *head pose*, the annotator checked whether the sighted passerby was looking at the blind user. As for *interaction*, the annotator marked whether the blind or the sighted participant was speaking while consulting the video recordings from the stationary camera. Annotations for *distance* and *position* were provided by two researchers, who first annotated independently and then resolved disagreements. The initial annotations from the two researchers achieved a Cohen's kappa score of 0.89 and 0.90 for distance and position, respectively. *Distance* was annotated as $<$ *6ft.*: within 6 feet (2 meters), *near*: farther than 6 feet (2 meters) but within 16 feet (5 meters); *far*: farther than 16 feet (5 meters). Distance in frames where the passerby was not included was annotated as

n/a. As for *position*, the annotators checked whether the person was on the *left*, *middle*, or *right* of the frame. This relative position was determined based on the location of the center of the person's face within the camera frame. When the face was not included, the center of included body part was used. The annotators resolved all annotation disagreements together at the end.

7.2.2.4 Data analysis

For our *quantitative analysis*, we mainly used descriptive analysis on our annotated data to see what visual information about passersby blind participants captured using the smartglasses and how well our system performed to extract proxemic signals (*i.e.*, the presence, position, distance, and head pose of passersby). We adopted the common performance metrics for machine learning models (*i.e.*, F1 score, precision, and recall) to measure the performance of our system in terms of extracting proxemic signals.

For our *qualitative analysis*, we first transcribed blind participants' responses to the open-ended interview questions. Then, we conducted a thematic analysis [229] on the open-ended interview transcripts. One researcher first went through the interview data to formulate emerging themes in a codebook. Then, two researchers iteratively refined the codebook together while applying codes to the transcripts and addressing disagreements. The final codebook had a 3-level hierarchical structure—*i.e.*, level-1: 11 themes; level-2: 20 themes; and level-3: 71 themes.

7.3 Results

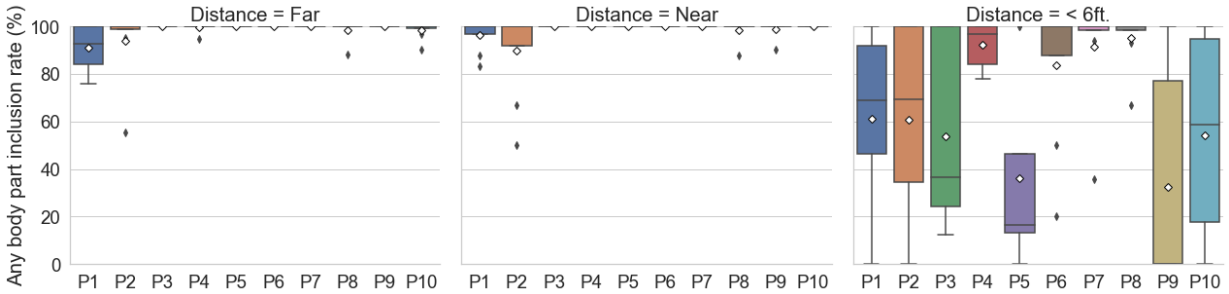
Although our data collection scenario is somewhat limited, the images, videos, and logs collected from 80 pairs of blind and sighted individuals allow us to account for variations within

a single blind participant and across multiple blind participants. Our observations, contextualized with blind participants' feedback, provide rich insights on the potential and limitations of head-worn cameras.

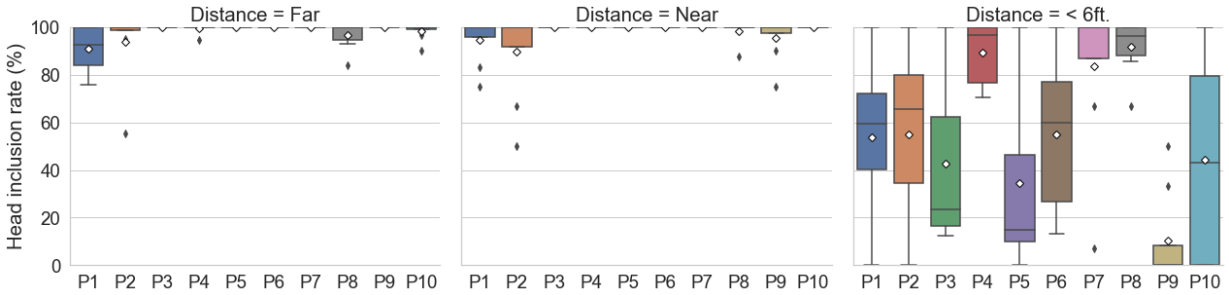
7.3.1 Visual Information Accessed by a Head-worn Camera

In a real-world setting, where both blind users and other nearby people are constantly in motion, it is important to understand what is being captured by assistive wearable cameras—smartglasses in our case. What is visible dictates what kind of visual information regarding passersby's proxemics can be extracted. In this section, results are based on our analysis of 3,175 camera frames from the head-worn camera and those ground-truth annotations.

At what distance is a passerby captured by the the camera? Figure 7.4 presents that a relatively high percentage of camera frames captured a passerby at a distance larger than 6 feet in a corridor that was 79 inches wide; this varied across blind participants with some outlier cases being as low as 50%. Within 6 feet, blind participants seemed to have different tendencies of including either the face or any body part of passersby, which average inclusion rates tend to be lower than those when passersby were more than 6 feet away. When sighted passersby were more than 16 feet away (*far*), part of their body and head were included on average in 98% (SD = 3.5) and 98% (SD = 3.6) of the camera frames, respectively. When sighted passersby were less than 16 feet but more than 6 feet away (*near*), the inclusion ratios of any body part and head were 98% (SD = 3.5) and 98% (SD = 4.3), respectively. However, when passersby were within 6 feet (2 meters), there was a quick drop in the inclusion rate with high variability across the dyads—passersby's head and any body part were included on average in 56% (SD = 39.1) and



(a) Inclusion of any body part of passersby in camera frames per distance category.



(b) Inclusion of the head of passersby in camera frames per distance category.

Figure 7.4: Diagrams presenting inclusion rates of any body part or head of passersby in camera frames per distance. A relatively high percentage of camera frames capture a passerby (face and body) at a distance larger than 6 feet in a corridor that is 79 inches wide; this can vary across blind participants with some outlier cases being as low as 50%. Within 6 feet it can get unpredictable with faces often being included less.

66% (SD = 39.6) of the camera frames, respectively. For example, the camera of P9, who tended to walk fast, consistently did not capture the passerby at this distance.

Beyond fast pace, there are other characteristics that can contribute to low inclusion rates such as veering [64] and scanning head movements. We observed little veering, perhaps due to hallway acoustics. However, we observed that many blind participants tended to move their head side to side while walking and more so while interacting with sighted participants to ask them about the nearby office number. This behavior in combination with the low sampling rate of one frame per second seemed to contribute to exclusion of the passersby from the camera frames even when they were in front or next to the blind participants. These findings suggest a departure from naive fixed-rate time-based mechanism (*i.e.* 1 frame/sec) to more efficient dynamic sampling (*e.g.* increasing frames on proximity) as well as fields of view and computer vision models that can account for such movements.

What proportion of a passerby is being captured by the camera while approaching versus while interacting? The types of visual information and proxemic signals desired by a blind user can differ by proximity to a passerby and whether that person is merely approaching or interacting with them. Nonetheless, what can be accessible will depend on what part of the passersby body is captured by the camera (*e.g.*, head, torso, and hands). In our dataset, we identified 11 distinct patterns. For example, a *(Head, Torso)* pattern indicates that only the head and torso of a passerby were visible in a camera frame. Figure 7.5 shows the average rate for each pattern and whether a passerby was *approaching* or *interacting* with a blind participant. It also includes a break-down of rates across participants.

When passersby were more than 16 feet away (*far*) and approaching, blind participants tended to capture the passersby in their entirety in more than 81% of the overall frames. This

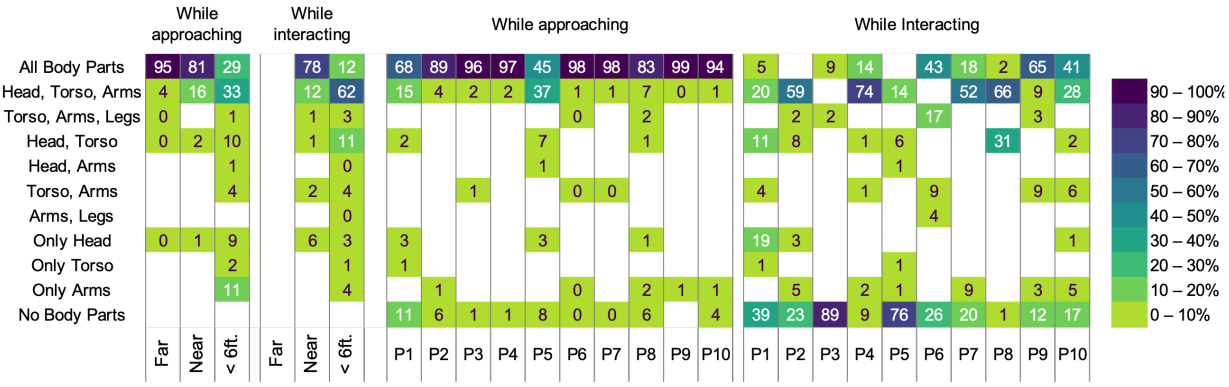


Figure 7.5: Heatmap diagram presenting the rates of body inclusion patterns in frames across distances, interactions, and participants.

number droped to 29% in the distance within 6 feet, making it difficult to estimate some of their proxemics, but the head presence remained high at 82%. These seemed to be consistent among the participants, except P1 and P5 whose rates were not only lower for including passersby in their entirety (68% and 45%, respectively) but also higher for not including them at all (11% and 8%, respectively).

We also analyzed the frames collected during blind participants' interactions with sighted passersby, typically within 16 feet (5 meters). Blind participants tended to capture the passersby in their entirety in about 78% of the overall frames for distances between 6 and 16 feet (2–5 meters). Closer than 6 feet, the number quickly dropped to 12%. More importantly, we observed a higher diversity in the inclusion rate across blind participants for what their camera captured when they were interacting with a passersby within 16 feet. Indeed, the breakdown of rates shows that all except P8, who was legally blind, and P4, who reported having a light perception and onset age of 49, did not include the passersby or their head in more than a quarter of their camera frames. For P3, P5, and P6, who had no prior experience with wearable cameras and reported onset age at birth, these rates were even higher reaching up to 91%. When inspecting their

videos from the static camera, we find that these participants were indeed next to the passersby interacting with them but they were not facing them.

7.3.2 Errors in Estimating Passersby Proxemic Signals

We evaluate how well models that rely on face detection can estimate passersby' proxemic signals (*i.e.*, *presence*, *position*, *distance*, and *head pose*) in camera frames from blind participants. Specifically, using visual estimates and those ground-truth annotations, we report results in terms of precision, recall, and F1-score (Table 7.3). These results capture blind participants' experience with the testbed, and thus provide context for interpreting their responses related to machine learning errors.

Table 7.3 presents that, when the presence of a passerby was detected, there was actually a passerby (precision=1.0)[7]. However, the model's ability to find all the camera frames where a passerby was present is low (recall=0.33)—even when their head was visible in the camera frame (recall=0.34). Our visual inspection on these false negatives informed that, for the majority (80%), this was due to the person being too far ($>$ 16 feet); thus, their face was too small for the model to detect, raising questions such as *How far away should a system detect people?*

Since the estimation of *position*, *distance*, *head pose* depended on the *presence* detection, we reported metrics on both (i) frames where passersby were present and (ii) frames where they were first detected. Once passersby were detected, the estimation of their relative position worked relatively well (precision=0.90; recall=0.84). However, the head pose estimation (*i.e.*, whether the passerby is looking at the user) was not as accurate (precision=0.77; recall=0.77). The dis-

[7]Such a high precision is not surprising since this detection task in our study was an imbalanced classification problem: there were two classes—*presence* and *no presence*—with the *presence* representing the overwhelming majority of the data points since there was always a passerby coming towards the blind participants. More so, there were no visual elements in the study environment that could confuse the model (*e.g.*, photos of people on the walls).

Table 7.3: Performance of our prototype system on the camera frames from blind participants.

Attribute	Comparison	Precision	Recall	F1-score
Presence	across all frames where a body part was present	1.0	0.33	0.49
	across all frames where head was present	1.0	0.34	0.51
Position	across all frames where someone was present	0.90	0.30	0.45
	across all frames where someone was detected	0.90	0.84	0.87
Distance	across all frames where someone was present	0.62	0.25	0.35
	across all frames where someone was detected	0.62	0.66	0.58
Head pose	across all frames where someone was detected	0.77	0.77	0.77

tance estimation was also challenging (precision=0.62; recall=0.66). This might be partially due to the limited training data, both in size and in diversity. A passerby's head size in an image, a proxy for estimating the distance, can depend on the blind user's height and camera angle.

7.3.3 Analysis on Qualitative Feedback

Accessing proxemics. The majority of blind participants (seven) appreciated having access to proxemics about nearby people through the smartglasses. P4 highlighted, *"I liked having the [proxemic] information. That means a lot when I know that there's a person, or how close they are to you or where they are[.] It's a good thing to be able to know who's there and that there's somebody there. That's important."* Blind participants also shared certain situations where they can benefit from such assistive cameras (*e.g.*, when nearby people do not provide enough non-visual cues). P2 highlighted, *"People are sometimes quite, and [I] don't know if they are approaching."* Some participants remarked that having access to proxemics related to a near passerby can help them make their decision proactively. For example, P7 mentioned, *"With the distance estimations, it gives [me] a couple of cues within range, where [I] can make a decision 'Do I want to do something or not?' "*

Participants also envisioned how these visual estimations could enhance their social engagement. For example, P1 said, *"It was novel [and] interesting to have it tell me things that I normally wouldn't even be paying attention to[.] I think it would be a great social interaction thing,"* and P4, *"It means a lot to be able to know who it is [and] who's there—that's a big deal to me for social interaction."* Some participants see its potential in helping them predict the outcome of their attempted social interaction by saying, *"If you know somebody is in front of you and looking [at you] or you know [where] to look (left or right), you can probably get their attention more easily"* (P5), and *"[GlAccess] certainly made it easy to know, when there was a potential [social interaction]—you didn't talk to the air[.] I would say that I like the ability to possibly predict the outcome of an attempted interaction"* (P7). They also envisioned using proxemics detection and person recognition to initiate their communication. P6 said, *"Walk down the hall and be able to greet a person by name when they hadn't spoken to me; that would be cool,"* and P7, *"Because [I] know that there is at least a person and how far away they are, [I'm] not simply chattering with air, which just occurred to me a couple times [in my daily life]."*

On the other hand, our analysis also reveals potential reasons why blind participants might not pay much attention to visual estimates. Since such visual information was mostly inaccessible in their daily lives, they did not rely on the visual information and also was unsure how to utilize them. More specifically, P7 and P9 remarked, *"[Getting visual information] never occurred to me. [Not sure if] I would want or need anything more than just listening to the voices around me and hearing the walking around me because I've never had that experience,"* and *"It's visual feedback that I'm not used to, I never functioned with, so I don't know, I'm not used to processing or analyzing things in a visual way. I love details, but I just never often think about them much related to the people,"* respectively. To address this issue, participants suggested providing train-

ing that helps blind users interpret such visual estimates as P7 said, *"I would need maybe to be trained on what to do [with the visual feedback]."*

Camera form factor. Six participants commented on how easy it was to have the smart-glasses on and access information, saying *"Just wearing the glasses and knowing that would give you feedback"* (P2); *"Just have to wear it and listen"* (P4); and *"Once [I] put on and just walk"* (P8). Some supported their statements by proving a rationale such as the head being a natural way to aim towards areas of interest and the fact that they didn't have to do any manipulations. For example, P4 and P9 commented, *"If it's just straight, then [I'm] gonna see what's straight in front of [me]. That's a good benefit[.] Seeing what directly in front of [me] is a bonus"* and *"I could turn my head to investigate or find out what's to my right or my left—that part was kind of nice,"* respectively. P7 and P10 commented on the manipulation, saying *"You don't have a lot of controlling and manipulating to do"* and *"And [it] doesn't take any manipulation—it just feeds new information. And that was easy. It took no interaction manually changed,"* respectively.

However, anticipating that some control would be eventually needed, P8 tried to envision how one could access the control panel for the glasses and suggested them being paired with the phone or support voice commands by stating *"Altering [it] would be an interface with your phone so that I could [control] it using the voice over on my phone, or using Siri to do it. The voice assistant [would] let you control."* Only one participant (P1) suggested deploying similar functionalities into one's phone *"And just use your rear facing camera and have it as a wearable, a pouch or something like that. Then you just flip that program on [the phone] when you want to use it."*

Other suggestions focused on the sound. At the time, Vuzix Blade did not have a built-in speaker, and their shape conflicted with bone conduction headphones. Four participants quickly

pointed out this limitation with statements such as *"Bone conduction [would] seem to work [with GlAccess] really good"* (P4) and *"[Need to keep] your ear canals open there for being able to hear traffic or whatever"* (P8). P4 suggested extending the field of view for capturing more visual information, stating *"I wonder if the field of vision for the camera could go a little bit wider because then it would catch somebody coming around the side more[.]"*

Interactions with the system. Four participants shared concerns on outside noises blocking out the system's audio feedback, or vice versa. P4 mentioned, *"The only time it could be harder [to listen to the system feedback] is in a situation with lots of noise or something. If it's very noisy, then it's going to be hard to hear,"* and *"There's a limit to how much you can process because you're walking. When we're walking[,] we're doing more than just listening to the [the feedback.] We listen to [others]. Listening is big [to us]."* Although GlAccess reported only feedback changed between two contiguous frames, this is possibly why they wanted to control the verbosity of the system feedback. P7 said, *"It was reacting to you every couple seconds or literally every second. Until I'm looking for something specific, it [is] a little overwhelming to have all of it coming in at the same time."*

Moreover, blind participants emphasized that they would need to access such proxemic signals in real time, especially when they are walking. They mentioned being misguided by stale estimations caused by network latency at times—P9 commented on this latency issue, *"I know about the latency in it. Because when [a person] and I were talking, I knew [the person] walked away, [but GlAccess] still reported that [the person was] in front of me."* To minimize the effects of the latency, some reported that they made quick adaptation, such as *"walking slow[ly]"* (P1) or *"stop[ping] and paus[ing]"* (P10), during the study. Nonetheless, the latency issue would make the use of such assistive wearable cameras impractical in the real world, as P8 put, *"If I'm*

walking at a normal pace [and] it's telling me somebody is near, I'm probably going to be 10 to 15 feet past [due to the latency]."

Participants also suggested providing options for customizing the feedback delivery. For example, P7 wanted the system to detect moments for pausing or muting the feedback delivery, saying *"Once a conversation is going on, it's not reporting everyone else around; [otherwise, it's] distracting [me]."* Also, four blind participants commented on verbosity control, which we explored a bit in our study by implementing two ways of communicating the distance estimation. P2 mentioned, *"I like that you had options [for verbosity], and [I] could choose that."* At the same time, we observe that blind people may have different definitions of spatial perception—*i.e.*, the definition of *far* and *near*. P4 remarked on this, *" 'Far'–'near' thing wasn't so good. I liked hearing the the number of feet for the distance or my concept of far is different than your concept of far and near[.]"*

Estimates and errors. Seven blind participants mentioned being confident of interpreting person recognition output. It seems that they gained their trust in the person recognition feature since they were able to confirm the recognition result with a person's voice. P2 said, *"I learned [a lab member]'s voice, and when it said '[lab member]', it was her; or at least I thought it was."* Similar to any other machine learning models, however, our prototype system also suffered from errors such as false negatives (Section 7.3.2). Participants seemed to be familiar with technical devices making errors and understood that it is part of technology improvement, as P1 commented, *"If it didn't work, then I figured that's just part of the the system growing, so that's okay. It's just part of learning how to use it."* To prevent from making their decision based on errors, six participants mentioned aggregating multiple estimates from the system, instead of relying on a singular estimate, after realizing the possibility of errors in visual estimates. P9 highlighted, *"If*

[the] information was not what I felt either incorrect or not certain, I would wait and give the system a chance to focus better or to process it differently."

Even with errors in visual estimates, blind participants said that it can help them know that *"there was somebody [at least]"* (P6). On the other hand, seven participants pointed out that there would be no guaranteed way for them to verify the visual estimates—*i.e.*, detecting errors in visual estimates is inaccessible. For example, P3 did not know if there was even an error because *"[P3] can't really recall that where it wasn't right."* Participants who even experienced something odd mentioned that it was difficult for them check whether that was an error, or not. P4 commented on this issue by saying, *"One time [GlAccess] goofed, I think, but I don't know if that [was an error.] I don't know if there were more people in the hallway."* Due to the inaccessibility of feedback verification, P10 mentioned being dependent on the system feedback, *"Unless someone was there saying that listening to same thing I was listening to and tell me if it was right or wrong, I had to depend on it."* We suspect that these experiences led them to expecting future assistive systems to provide more reliable visual estimates, as P3 commented, *"[Future glasses need to] reassure that it's accurate in telling you like distance [how far] people are."*

Teachable person recognition. Although blind participants did not experience personalizing their smartglasses in the study, they envisioned that such assistive technologies would be able to recognize their family and friends once they provide their photos to the device. For instance, P1 expected this teachable component by saying, *"[Person recognition] means [that] I have to create a database of pictures. And I wouldn't mind doing that if my friends wanted to."* Recognizing our lab members during the study seemed to help them see the benefits of recognizing their known people via smartglasses. P6 was excited about using this recognition feature to greet her

family and friends by their name even *"when they hadn't spoken to [P6]."* Also, P10 mentioned training smartglasses to find, *"out of a group, a person that [P10 is] looking for."* P1 commented on sharing photos with other blind people to recognize each other through such smartglasses, saying *"When I was at the American Council of the Blind convention, there were rooms full of blind people, [and] none of [us] knew who all was in the room. So wouldn't it be cool if we all had each other's pictures [to teach smartglasses about us]?"*

7.4 Discussion

Our exploratory findings from the user study lead to several implications for the design of future assistive systems supporting pedestrian detection, social interactions, and social distancing. We discuss those in this section to stimulate future work on assistive wearable cameras for blind people.

7.4.1 Implications

Accessing proxemic signals using head-mounted cameras. Section 7.3.1 reveals that some camera frames from blind participants did not capture passersby's face but other body parts (*i.e.*, torso, arms, or legs), especially when they were near the passersby. Although some visual cues such as head pose and facial expressions [11] cannot be extracted from the other body parts, detecting those may help increase accessibility of someone's presence [230, 231]. Also, we see that proximity sensors can enhance the detection of a person's presence and distance as those sensors can provide depth information. In particular, such detection can help blind users to practice social distancing, which has posed unique challenges to people with visual

Table 7.4: Information about non-handheld cameras designed for helping users with visual impairments access visual surroundings.

Camera Device	FOV	Camera Placement	Design Purpose
Horizon smartglasses	120-degree horizontal	head-mounted	Aira (agent-support) [24]
OrCam MyEye	*not available*	head-mounted	face recognition & color identification [10]
Microsoft HoloLens1	67-degree horizontal	head-mounted	person detection [61], stair navigation [233]
Ricoh Theta	190-degree horizontal[a]	head-mounted	person detection [61]
ZEDTM2K Stereo	90-degree horizontal	suitecase-mounted	person detection [215]
Vuzix Blade	64-degree horizontal	head-mounted	person detection [171]
RealSense D435	70-degree horizontal	suitcase-mounted	person detection [234][b]

[a]It creates a 360-degree panorama by stitching multiple photos.
[b]Two cameras were placed side by side to obtain 135-degree horizontal POV.

impairments [200]. The sensor data however needs to be processed along with person detection, which often requires RGB data from cameras. Future work that focuses on more robust presence estimation for blind people may consider incorporating proximity sensors with cameras.

Regarding delivery of proxemic signals, it is important to note that blind people may have different spatial perception [232] and could thus interpret such proxemics feedback differently, especially if the feedback was provided with adjectives related to spatial perception, such as *'far'* or *'near'* (Section 7.3.3). Other types of audio feedback (*e.g.*, sonification) or sensory stimulus (*e.g.*, haptic) may overcome the limitations that the speech feedback has. On the other hand, blind participants' experiences with our testbed may have been affected by its performance and thus could differ if an assistive system had more accurate performance or employed different types of feedback. Future work should consider investigating such a potential correlation.

Various factors affecting blind people's behaviors. Several factors, such as blind people's tendency of veering [64] and behavior of scanning the environment when using smartglasses (Section 7.3.1), can change camera input on wearable cameras and thus may affect the performance of detecting proxemic signals. From our study, we observed that the head movement of blind people often led to excluding a passerby from camera frames of a head-worn camera. Al-

though we did not observe blind participants' veering tendency in our study, the veering, which can vary by blind people [64], could change their camera aiming and consequently affect what is being captured. Also, as discussed in earlier work [232], blind people's onset age affects the development of their spatial perception. Observing that our blind participants had different onset ages, we suspect that some differences in their spatial perception may explain their different scanning behaviors with smartglasses. Future work should investigate these topics further to enrich our community's knowledge on these matters.

Moreover, using a wide-angle camera may help assistive systems to capture more visual information (*e.g.*, a passerby coming from the side of a user) and tolerate some variations in blind people's camera aiming with wearable cameras. The smartglasses in our user study, having only a front view with the limited horizontal range (64-degree angle), are unable to capture and detect people coming from the side or back of a blind user. As described in Table 7.4, some prior work has employed a wide-angle camera [61] or two cameras side by side [234] to get more visual data, especially for person detection. However, accessibility researchers and designers need to investigate which information requires more attention and how it should be delivered to blind users. Otherwise, blind users may be overwhelmed and often distracted by too much feedback (Section 7.3.3).

In addition, aiming with a wearable camera can vary by where the camera is positioned (*e.g.*, a chest-mounted camera) as earlier work on wearable cameras for sighted people indicates that the position of a camera can lead to different visual characteristics of images [63]. In the context of person detection, prior work investigated assistive cameras for blind people in two different positions (*i.e.*, either on a suitcase [215, 234] or on a user's head [61, 171]). However, little work explores how camera positions can change the camera input while also affecting blind

users' mental mapping and computer vision models' performances. We believe that it is important to study how blind people interact with cameras placed in different positions to explore design choices for assistive wearable cameras.

Errors and different feedback modalities. Data-driven systems (*e.g.*, machine learning models) inherently possess the possibility of generating errors due to several factors, such as data shifts [235] and model fragility [236]. Since the system that study participants experienced merely served as a testbed, it was naturally error prone (Section 7.3.2), which is largely true for any AI system, and consequently affected blind participants' experiences. (Section 7.3.3). In particular, blind participants pointed out inaccessibility of error detection in visual estimates. To help blind users interpret feedback from such data-driven assistive systems effectively, it is imperative to tackle this issue (*i.e.*, *How to make such errors accessible to blind users?*). One simple solution may be to provide the confidence score of an estimation, but further research is needed to understand how blind people interpret this additional information (*i.e.*, confidence score) and to design effective methods of delivering such information.

Furthermore, different feedback modalities are worth exploring. In this work, we provided visual estimates in speech, but blind participants sometimes found it too verbose and imagined that it would be difficult for them to pay attention to speech feedback in certain situations, *e.g.* when they are in noisy areas or talking to someone. To prevent their auditory sensory system from being overloaded, other sensory stimuli such as haptic can be employed to provide feedback [237, 238]. It is however important to understand what information each feedback modality can deliver to blind users and how effectively they would perceive the information.

Potential interaction modalities and possible applications for smartglasses. Blind participants envisioned personalizing their smartglasses to recognize their family members and

friends (Section 7.3.3). To provide this experience for users, it would be worth exploring teachable interfaces, where users provide their own data to teach their systems [3]. Our community has studied data-driven approaches in several applications (*e.g.*, face recognition [11], object recognition [2,4,160], and sound recognition [239]) to investigate the potential and implications of using machine learning in assistive technologies. However, little work explores how blind users interact with teachable interfaces, especially for person recognition. Our future work will focus on this to learn opportunities and challenges of employing teachable interfaces in assistive technologies.

Moreover, in Section 7.3.3, blind participants suggested enabling future smartglasses to detect specific moments (*e.g.*, sitting in a table or talking with someone) to control feedback delivery. We believe that egocentric activity recognition [99, 240, 241], a widely-studied problem in the computer vision community, can help realize this experience. However, further investigation is required for adopting such models in assistive technologies, especially for blind users. For instance, one recent work proposed an egocentric activity recognition model assuming that input data inherently contains the user's gaze motion [241]. However, that model may not work on data collected by blind people since the assumption, (*i.e.*, eye gaze), only applies to a certain population (*i.e.*, sighted people).

7.4.2 Limitations

We simplified our user study to have only one passerby in a corridor since our prototype system was able to detect only one person at a time. In the future, designers should consider handling a case where more than one passerby appears. When more than one passerby is included in a camera frame, a system needs to detect a person of the most interest to blind users and then

delivering the person's proxemics to the users. Such a method should help the users perceive the output, efficiently and quickly. For example, detecting a person within a specific distance range could be a naive but straightforward way.

There are still challenges related to hardware, privacy, and public acceptance, which keep wearable cameras from being employed for assistive technologies. For example, Aira ended their support for Horizon glasses in March 2020, due to hardware limitations [242]. Also, prior work investigated blind users' privacy concerns over capturing their environments without being given access to what is being captured [28, 243], and social acceptance of assistive wearable cameras as the always-on cameras can capture other people without notification [16, 82, 171].

Beyond the high cost of real-time support from sighted people (*e.g.*, Aira), as of now, there is no effective solution that can estimate the distance between blind users and other people in real time. Computer vision has the potential to provide this proxemic signal in real time on a users' device. No recordings [171] and privacy-preserving face recognition [244, 245] may help address some privacy issues, but more concerns may arise if assistive wearable cameras need to store visual data (even on the device itself) to recognize the users' acquaintances. Future work should consider investigating associated privacy concerns and societal issues, as these factors can affect design choices for assistive wearable cameras.

7.5 Summary

In this chapter, we explored the potential and implications of using a head-worn camera to help blind people access proxemics, such as the presence, distance, position, and head pose, of pedestrians. We built GlAccess that focused on detecting the head of a passerby to estimate

proxemic signals. We collected and annotated camera frames and usage logs from ten blind participants, who walked in a corridor while wearing the smartglasses for pedestrian detection and interacting with a total of 80 passersby. Our analysis results show that their smartglasses tended to capture passersby's head when they were more than 6 feet (2 meters) away from passersby. However, the rate of including the head dropped quickly as blind participants were getting closer to passersby, where we observed variations in their scanning behaviors (*i.e.*, head movements). Blind participants shared that using smartglasses for pedestrian detection was easy since they did not have to manipulate camera aiming. Their qualitative feedback also led to several design implications for future assistive cameras to consider. In particular, we found that errors in visual estimates need to be accessible to blind users to better interpret such outputs for their autonomy.

Chapter 8: Social Acceptance of Pedestrian Detection System on Head-Worn Camera[1]

With advances in machine learning and computing devices that are small and light enough to be worn on one's body, wearable cameras hold the promise to increase real-world accessibility for people with visual impairments (*e.g.* [60,61,79,246]) with few ideas making it to commercial products such as OrCam [10], Aira [24] and eSight [81]. Given that these devices can capture information in the surroundings that people are uncomfortable sharing or that could be used for surveillance, public use of wearable cameras is conditional on societal consent. Thus, beyond technical challenges, such as battery life, accuracy, form factor, and split-attention, issues such as social acceptability and privacy concerns [16] hold back their use for more independent living.

With this chapter, we aim to contribute to the ongoing discussion on the social tensions with wearable cameras that have sensing capabilities for accessibility (*e.g.* [82,247,248]). Specifically, we focus on pedestrian detection with wearable cameras for blind users. While this technology can provide equitable access to approaching pedestrians with implications for ensuring one's safety [249] or managing social interactions [250], its always-on nature raises privacy concerns for parties that may or may not realize that they are being recorded and/or visually interpreted.

The main contribution of this work is that we explore this tension from both perspectives,

[1]This chapter presents joint work with Professors Hernisa Kacorri and Chieko Asakawa, Dr. Daisuke Sato, and Saki Asakawa published in ACM CHI 2020 [171].

(a) Passerby perspective video: A recording from the perspective of pedestrians watching a blind person who wears either smartglasses or a GoPro camera with a head strap, approaches, and walks by them.

(b) User perspective video: A recording showing a blind and a sighted person walking toward each other in a corridor. When in proximity, the perspective shifts to the wearable camera showing what it sees and what the blind user hears about the passerby behind the scenes.

Figure 8.1: Video stimuli shown to our sighted participants during the online and in-situ studies. Both perspectives were included in the online study; in-situ only the user perspective.

those of sighted passersby and blind users, taking into account camera visibility and remote versus in-person experience with a working prototype; dimensions lend an opportunity for further insights into prior findings from Profita *et al.* [16] and Ahmed *et al.* [82]. More specifically, our first study is a balanced online 2x2 factorial-design experiment. Participants (N=206), recruited through Amazon Mechanical Turk, responded to technology attitude and other questions before and after being exposed to videos portraying passerby and user perspectives. Videos varied in terms of the wearable camera (visible head-mounted GoPro versus 'invisible' camera in smartglasses) and in terms of the blind actor (male adult with a white cane versus female older adult with a guide dog), as shown in Figure 8.1. Our second study is an in-person experience

study, where blind participants (N=10) wear a working prototype for pedestrian detection and pass by sighted participants (N=40). Since sighted participants experience a scenario similar to the passerby perspective video, they are only shown the user perspective video. Duerden and Witt [251] indicates that direct experience catalyzes people's perception of technology more than remote experience. We explore how this plays when trying to capture people's perceptions toward assistive wearable cameras for pedestrian detection designed for blind people.

We find that our blind and sighted participants held differing opinions on privacy concerns brought forth by assistive wearable cameras. They displayed a conflict of interest regarding which visual features should be detected by an assistive wearable camera; *e.g.*, blind participants favored estimations of *head pose* and *facial expression* of pedestrians, but sighted passerby participants did not want such an assistive system to detect these features. We observed that sighted participants changed their opinions by being more reserved on assistive tech with wearable cameras after learning what information it can glean from pedestrians. We noticed that the way bystanders experienced this technology (personally or remotely) seems to relate with their opinions; with in-person participants being more negative though this could be explained by other factors and requires a follow up study to confirm directionality. Our findings suggest that it is important to understand the conflict of interests between users and bystanders regarding such technology while also confirming that people are amicable toward assistive uses of wearable camera technology. Our results suggest that addressing the conflict in design choices between potential users and pedestrians would help increase social acceptance of it. We recommend that other factors, such as the way of exposure to technology, and the awareness of the data being captured and processed by the technology, need to be considered to obtain a clearer understanding of people's perceptions and attitudes.

8.1 Motivation

Our research is built upon prior work concerning the social acceptance of wearable technology as well as face detection and recognition technology for people with visual impairments. We discuss prior work to understand people's perceptions of wearable technology and learn about assistive camera systems designed for people with visual impairments.

Wearable technology that expects user input often uses gestural interfaces. As gestural interactions between users and wearable devices can cause social tensions in certain situations, researchers have explored the social acceptance of such technology [15, 26, 27, 252–254]. This social acceptance evaluation focuses primarily on the perspective of users (their perceptions of the technology and their comfort with using such gestures) [26, 27, 252], since users are the only source of input for such wearable devices. On the other hand, there is prior work evaluating social acceptance of gesture-based wearable technology from the bystander's perspective [15, 254]. Videos presenting the uses of wearable devices are often provided to participants so that participants can indirectly experience the technology and then complete a questionnaire for social acceptance.

Understanding the social implications of wearable technology from a bystander's perspective becomes more popular when concerning wearable camera devices [16, 80, 255–259], as bystanders are often involved with the technology more explicitly than before. Although wearable camera technology, such as Google Glass, has caused epidemic controversial privacy concerns, prior work shows that people are more positive toward wearable camera technology when aware that such systems are designed for assistive purposes [16]. Participants in the prior study remotely experienced the technology by viewing videos recorded from the bystander's viewpoint,

yet the authors also acknowledge that in-person experiences with the technology (*e.g.*, seeing or interacting with users wearing such cameras) may engender different social implications. In this chapter, we further investigate these unanswered questions regarding the impact of increased awareness of the data used by wearable cameras and in-person versus remote experiences of the cameras on social acceptance.

8.2 Methods

8.2.1 Online Study

We first explore social acceptance of wearable camera technology for blind people with online sighted participants. Amazon Mechanical Turk was used as it provides easy access to a large pool of people. More than 200 participants in the U.S. answered our questions before and after watching the passerby perspective video and the user perspective video (Figure 8.1)[2]. We created our online survey through SurveyGizmo and linked it to Amazon Mechanical Turk.

8.2.1.1 Participants

A total of 206 individuals (74 female, 132 male) from the U.S. were recruited[3] via Amazon Mechanical Turk. As in Profita *et al.* [16], we specified the USA-only constraint to minimize potential data variations due to cultural and regulatory differences. Participants ranged from 19 to 70 in age (μ=35.78, σ=9.9); only adults, age 18 or older, were allowed to participate in our online survey. Our online participants tended to be slightly positive toward broader technology

[2]The questionnaire and videos are available on `https://iamlabumd.github.io/chi2020_lee/`
[3]IRB Protocol Number is *STUDY2019_00000294*.

Table 8.1: Video characteristics and reference codes.

Camera	*Blind Actor*	*Assistance*	*Code*
GoPro	male	white cane	**GoPro+cane**
smartglasses	male	white cane	**Glasses+cane**
GoPro	female	guide dog	**GoPro+dog**
smartglasses	female	guide dog	**Glasses+dog**

($\mu = 1.63$, $\sigma = 0.66$)[4] and wearable technology ($\mu = 1.16$, $\sigma = 1.24$)[4].

8.2.1.2 Video Stimuli

Participants watched two videos that presented assistive wearable camera technology in two different perspectives, respectively: the passerby perspective and the user perspective. We created the videos using two different wearable devices—either smartglasses or a GoPro—and two different sets of actors—(a female pedestrian and a male blind user with his white cane) or (a male pedestrian and a female blind user with her guide dog). This totaled to four different versions of each video. Figure 8.1 describes the videos, and Table 8.1 shows the videos' characteristics and reference codes.

The passerby video presents a blind user wearing a camera device and walking through a corridor. This video shows how blind users wearing such a camera would be observed by pedestrians. On the other hand, the user video introduces four scenarios using the camera for assistive purposes, focusing more on the user's point of view. It presents how pedestrians would be captured by the camera as follows:

Scenario 1 The wearable camera detects the pedestrian's presence and estimates the distance between the user and pedestrian as well as the head pose of the pedestrian.

[4]Responses to attitudes questions are converted into -3 to 3 with -3 indicating the *most negative*, 0 *neutral*, and 3 the *most positive*.

Scenario 2 The wearable camera detects simple visual features of the pedestrian, such as age and gender, as well as the pedestrian's distance and head pose.

Scenario 3 The wearable camera not only estimates the distance and head pose but also detects more visual features, such as age, gender, ethnicity, and hair color.

Scenario 4 The wearable camera recognizes the pedestrian and estimates the distance and head pose.

8.2.1.3 Procedures

On the instructions page, we first mentioned that our online survey contained secret questions for checking participation and determining compensation to encourage participants to read and answer all the questions thoroughly; it was adapted from the performance-based payment approach [260]. After that, the online survey opened with demographic questions and several questions about attitudes toward technologies. Once participants completed this questionnaire, the survey platform randomly assigned a pair of videos showing a different set of actors and either of the wearable cameras. On the survey, we described that the devices assisted blind people.

There were two secret questions (one after the passerby video and another after the user video) that we used to check whether online participants watched the videos; we marked a response as invalid and did not use it for our analysis if a participant had a wrong answer to either of these questions. The secret questions asked about what the blind user was holding in the passerby video and what features the prototype system detected from the pedestrian in the user video.

After watching the passerby video, participants were asked to answer questions about social acceptance of assistive wearable camera technology. Then, they watched the user video and

were asked if they would be okay with each of the scenarios. Once they responded to the scenario questions, they were asked to answer some of the social acceptance questions, which were the same as the ones asked after the passerby video. These questions were intentionally asked again to study changes in the participants' perceptions of the technology after the user video that informed them of what sort of information the system collected from pedestrians and used for assistive purposes. The back button on the online survey was disabled to prevent participants from altering their answers to previous questions. Participants were compensated $2 for their participation as long as they completed the survey.

8.2.1.4 Questionnaire

A questionnaire about social acceptance toward assistive wearable camera technology consists of questions that are grouped into eight categories:

- **OkayAssistUse**: Okay to see blind people using the technology in public.
- **OkayAnyUse**: Okay to see any people using the technology in public.
- **PrivacyConcerns**: Privacy issues with blind people using the technology in public.
- **UncomfortableAssistUse**: Uncomfortable to see blind people using the technology in public.
- **OkayRecording**: Okay with the camera recording photos/videos for improvement.
- **OkaySingleUse**: Okay with the camera using photos/videos for one-time detection.
- **UncomfortableRecording**: Uncomfortable with the camera recording photos/videos.

- **UncomfortableSingleUse**: Uncomfortable with the camera using photos/videos for one-time detection.

Questions under four categories (*OkayAssistUse, OkayAnyUse, PrivacyConcerns, UncomfortableAssistUse*) were asked after the passerby video and the user video, respectively. We created questions based on prior questions asking about social acceptance of wearable technology and attitudes toward technology [132, 261]. For some aspects of social acceptance, we asked participants about their *acceptance* (*i.e.*, Okay-) and *discomfort* (*i.e.*, Uncomfortable-), as people might accept the technology but feel uncomfortable about it, or vice versa.

In addition, after the user video, participants answered questions about the four different uses of such assistive technology as well as one additional use (*Recognition of any people*). Following are the descriptions of the five questions:

- **Scenario 1 (presence, distance)**: Okay with the camera detecting my presence and distance between the user and me.

- **Scenario 2 (basic visual features)**: Okay with the camera detecting my basic visual attributes, such as age and gender.

- **Scenario 3 (more visual features)**: Okay with the camera detecting more of my visual attributes, such as age, gender, hair color, facial expression, ethnicity, and actions.

- **Scenario 4 (recognition)**: Okay with the camera recognizing me (*i.e.*, my name) when I know the blind user.

- **Recognition of any people**: Okay with the camera recognizing me (*i.e.*, my name) when I don't know the blind user.

Participants were asked to respond all the questions above with a 7-point Likert scale: *Strongly disagree, Moderately disagree, Slightly disagree, Neutral, Slightly agree, Moderately agree, Strongly agree*.

8.2.2 In-Situ Study

The goal of our in-situ study is to understand conflicts of interest in such assistive technology between blind users and pedestrians and to further evaluate their opinions on the technology by comparing results from our online survey.

8.2.2.1 Prototype System

We further wanted to know how people would perceive assistive wearable cameras if they were exposed to such technology in person. For this evaluation, we built a working prototype system using a wearable camera device to conduct an in-situ user study. The prototype system is designed to help blind people access visual information about pedestrians around them. We call our prototype system *GlAccess* (Glasses + Access), which estimates visual attributes of people who pass by our system users. *GlAccess* consists of a server and a client. The client is an app running on a pair of smartglasses that communicates with the server to obtain visual features of a pedestrian as described in Figure 8.2. This prototype system is the same as that described in Section 7.2.1.

Server: The server is responsible for detecting visual features of and recognizing a pedestrian captured in photos sent by the client. Once the server receives a photo from the client, it tries to estimate visual features—such as name, age, gender, and head pose—of a person in the

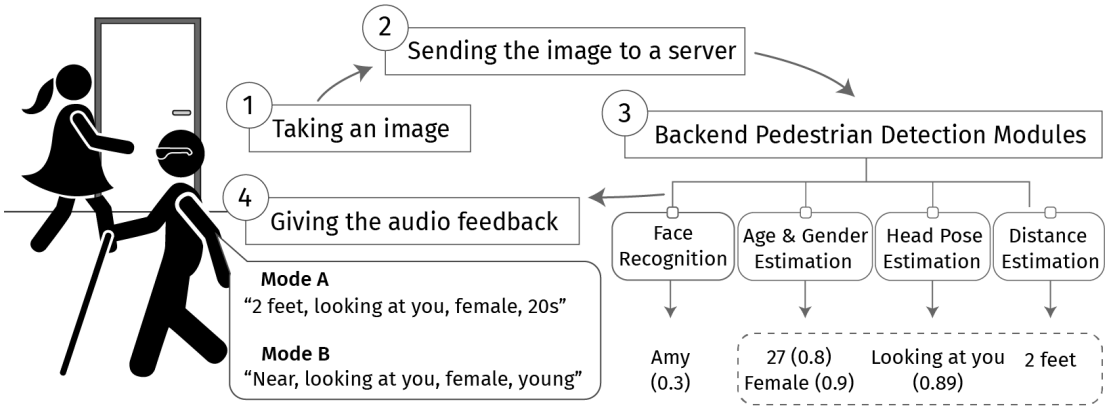


Figure 8.2: Diagram of data processing of *GlAccess*.

photo. Following is a list of capabilities of the server: face detection and recognition, and age, gender, head pose, and distance estimations.

Client: The client is responsible for delivering feedback to users based on the image analysis from the server. It continuously requests the server to analyze one image at a time. Once the system detects a pedestrian, it generates a text message and speaks it by using text-to-speech. The client remains silent if no one is detected. To prevent message delivery from being interrupted by new responses, the client will ignore the new message if it is in the middle of message delivery of a previous message.

The client has two different modes, *Mode A* and *Mode B*. The system composes the messages in order of distance, head pose, (name or gender & age), and position; *e.g.*, "15 feet, looking at you, male, 30s, on the left." The gender and age estimations are provided to users if the system is not confident of the name estimation (*i.e.*, face recognition). Each mode has a different way of creating a feedback message for distance and age; but, the other features (head pose, name, gender, position) are delivered same way in both Mode A and B. If the confidence score[5] of each estimation is below a certain threshold, the system does not deliver that estimation. Based on our

[5] $0 \leq$ confidence score ≤ 1; the higher, the more confident.

pilot testing, we empirically set two different thresholds: 0.5 for gender, age, and name estimations, and 0.7 for head pose estimation. Note that the age and head pose estimations are binary classifications, and the distance and position estimations do not have confidence scores as these estimations are proportionally computed with the size of a detected face.

Mode A: Quantitative feedback mode. It uses the foot metric for distance estimation and approximation for age estimation— *e.g.*, "2 feet, looking at you, female, 20s."

Mode B: Concise feedback mode. It simply says whether a pedestrian is detected to be *near* ($<$ 15 feet) or *far*. Age is estimated as *young* (age $<$ 35), *middle-age* (35 $\leq$ age $<$ 55), or *older*—*e.g.*, "Near, looking at you, female, young."

Implementation: The server application was written in Python with the Flask framework and run on a GPU-powered Ubuntu 18.04 Linux system. For face detection and alignment, we employed Insightface [262] built upon the Multi-task Cascaded Convolutional Networks (MTCNNs) [225] that detects faces. This model was composed of several models that not only detect face embedding (512-dimension data) but also estimate age and gender of a detected face. Using the face embedding data, we trained a SVM classifier to recognize five of our lab members; blind participants interacted with them during user study (see next section). Moreover, we fine-tuned[6] the face embedding model on a dataset for head pose estimation; the dataset included the head pose database [227] and our head pose images collected using the smartglasses. The head pose estimation model simply checked whether or not a pedestrian was looking at a user. The distance between a pedestrian and a user was estimated based on the height of a detected face. The distance estimation function was defined with pairs of the actual distance and the face height from

[6]The SGD optimizer was used with the following hyperparameters: learning rate=0.005, batch size=8, epochs=100.

Table 8.2: Demographic information about blind participants.

PID	*Gender*	*Age*	*Blindness*	*Light Persept.*	*Onset*	*Wearable Camera*
P1	F	71	Totally	N	63	Y
P2	M	66	Totally	N	birth	Y
P3	M	59	Totally	N	birth	N
P4	F	66	Totally	Y	17	N
P5	F	65	Totally	N	birth	N
P6	F	73	Totally	N	birth	N
P7	F	49	Totally	N	35	N
P8	M	60	Legally	Y	4	Y
P9	M	56	Totally	N	30	N
P10	M	71	Totally	Y	30	Y

our dataset. Since our user study was conducted in the U.S., we used *foot* for distance.

The client system was implemented as an Android app for smartglasses, Vuzix Blade [186]. The smartglasses ran on Android OS (5.1) and were connected to the server through Wi-Fi. Since the glasses did not have a speaker, a Bluetooth headset with earbuds was used so that users could hear audio messages; using a bone conduction headset was not feasible as its body conflicted with the glasses. During the development, we observed that it usually took less than one second to send a request to and receive a response from the server.

8.2.2.2 Participants

We recruited[3] a total of 50 participants: 10 blind participants and 40 sighted passerby participants.

Blind participants: Their ages ranged from 49 to 73 years (μ=63.6, σ=7.6). As described in Table 8.2, four (P1, P2, P8, P10) of them had prior experience with assistive wearable cameras, such as Aira [24], but many of them mentioned that they only used the technology a few times for trial purposes (a few times a month). Their responses to pre-study questions indicate that

our blind participants tended to be positive about general technology ($\mu = 2.02$, $\sigma = 0.42$)[4] and wearable technology ($\mu = 1.89$, $\sigma = 0.90$)[4]. Note that these blind participants are also described in Section 7.2.2.1.

Sighted participants: For each user study, we recruited four sighted passerby participants on our campus; as a result, there were 40 sighted passerby participants in total: 6 faculty/staff, 25 students, and 9 visitors. Their ages ranged from 19 to 39 years (μ=25.3, σ=4.8). Based on their responses to general attitude questions in the post-study sessions, we suspect that they were slightly positive about wearable technology ($\mu = 1.08$, $\sigma = 0.95$)[4]; there was no pre-study session for the sighted.

8.2.2.3 Study Procedures

There were four sessions (*pre-study*, *practice*, *main*, *post-study*) for blind participants and two sessions (*main*, *post-study*) for sighted passerby participants. In the main session, blind users and sighted passerby walked through a corridor and passed by each other as shown in Figure 8.3; there were one blind participant and one sighted passerby in the corridor at a time. Each blind participant passed by eight sighted passerby people: four people (two sighted participants and two of our lab members) for each mode evaluation (Mode A/B). In each mode evaluation, we also asked first two sighted people to look as they would normally and the other two people to look at their phones while walking through the corridor.

Blind participants: On average, the user study took two hours. Once blind participants arrived, they first met with our lab members who later participated in their user study as passerby so that the blind participants would recognize the lab members. Then, we began the pre-study

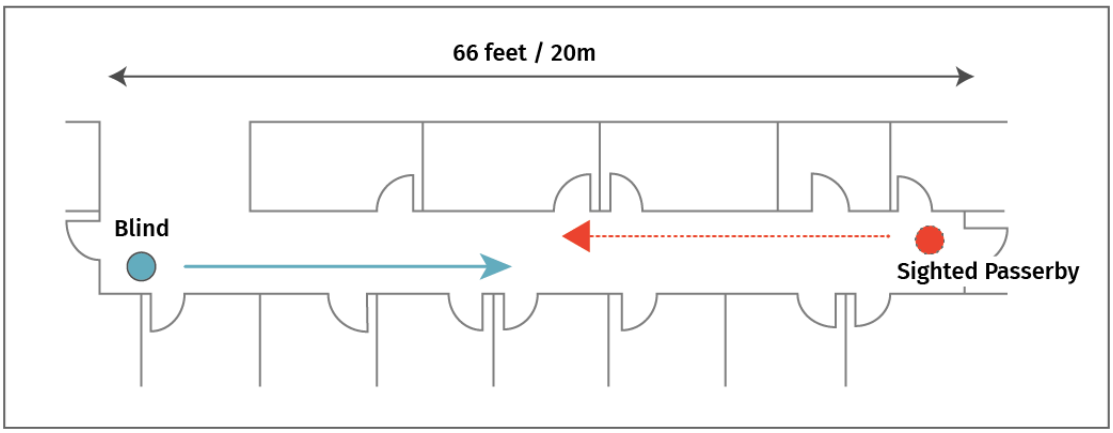


Figure 8.3: Starting points and walking directions of blind participants and sighted passerby participants in a 66-feet (20-meter) corridor. A blind participant wearing *GlAccess* and a sighted passerby participant walk toward each other.

Table 8.3: Categories and descriptions of social acceptance questions for blind participants, and mean (std.) of scores of their responses to the questions. Responses are converted to numeric values (-3 to 3): -3=*Strongly disagree*, 3=*Strongly agree*.

Category	*Description*	*Avg. Score*
OkayUseInPublic	Okay to wear the smartglasses in public	2.6 (0.7)
OkayAnyUse	Okay with any people using the technology in public	2.2 (1.0)
ComfortableUse	Feel comfortable using the smartglasses	2.5 (0.5)
FeelAccomplished	Feel more accomplished due to this wearable technology	1.6 (1.5)
HelpInteractWithKnown	Face recognition helps initiate social interactions with known people	2.1 (1.0)
HelpInteractWithUnknown	Feature detection helps initiate social interactions with unknown people	1.4 (1.6)
OkayRecording	Willing to provide many photos/videos of known people if they agree	2.1 (1.1)
PrivacyConcerns	Privacy issues with blind people using the technology in public	-1.0 (2.0)
DifficultFaceRecognition	Face recognition using a wearable camera is difficult	-2.0 (1.3)
DifficultFeatureDetection	Feature detection using a wearable camera is difficult	-2.0 (0.9)

session by obtaining their demographic information, attitudes toward general technology and wearable technology, and prior experience with wearable cameras. After that, the blind participants wore smartglasses and started using *GlAccess* to detect and recognize a person. During this practice, blind participants tried Mode A and Mode B to familiarize themselves with each of the modes. Once blind participants became familiar with our system, we started the main study session with sighted participants and our lab members. During this session, blind participants used *GlAccess* and were asked to perform the following instruction—"When the system detects that a person is nearby, ask the person about the closest office number." Using each feedback

mode, blind participants walked in the corridor four times to detect or recognize four different people, including two recruited passerby participants who were unknown to our system and two of our lab members who were known to our system, and performed the instruction given to them in the beginning of the session. After that, they answered questions about their experience and perceptions of the system. Table 8.3 describes ten categories of the questions.

Sighted passerby participants: Sighted passerby participants were asked to simply walk through the corridor. Before the study, we told them that there would be a camera recording when they walk in the corridor, but did not inform them of a wearable camera device. Once sighted participants finished walking in the corridor (*i.e.*, the walking activity), they completed our post-study questionnaire, which included the same questions of our online survey; but excluded the passerby video since they were personally engaged in this video's situation during the user study.

8.3 Results

8.3.1 Online Study

Among the total 206 responses, 22 incorrectly answered to our secret question(s) and were thus excluded from our analysis. As a result, we obtained 45, 47, 45, and 47 valid responses from participants who watched GoPro+cane, Glasses+cane, GoPro+dog, and Glasses+dog video pairs, respectively. Only the first 45 responses for each video pair were used for our analysis; hence, our analysis used 90 responses from each camera form (Glasses/GoPro). For analysis, we converted their responses from the 7-point Likert scale into numeric values (-3 to 3): -3 for *Strongly disagree*, 0 for *Neutral*, and 3 for *Strongly agree*.

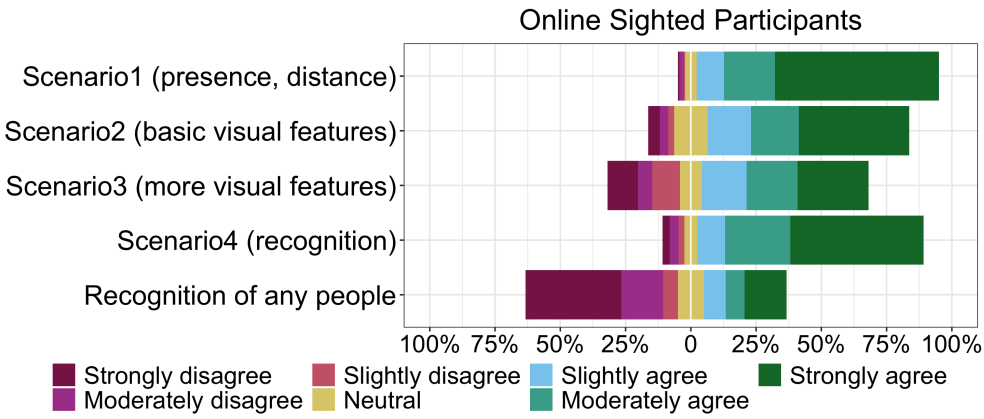
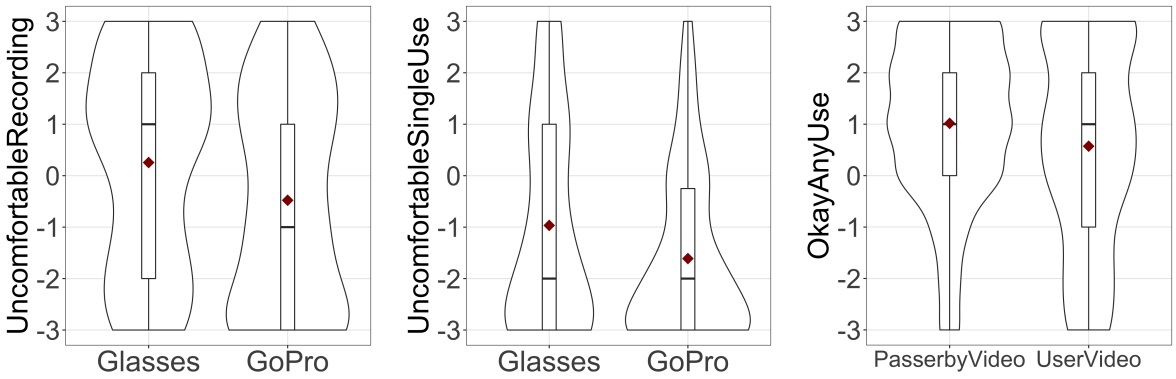


Figure 8.4: Online sighted participants' opinions on use cases of assistive wearable cameras. Participants were asked whether they would be okay with each of the use cases.

8.3.1.1 Perspectives of online sighted participants

Using all responses from the Glasses and GoPro groups, we first examined participants' perceptions of assistive wearable camera technology based on their responses to the social acceptance questions answered after the passerby video. In general, our participants tended to be positive toward the assistive wearable technology for all the eight categories. Online sighted participants tended to agree with the positive statements about the technology for assistive purposes, such as *OkayAssistUse* ($\mu = 2.2$, $\sigma = 1.1$), *OkayRecording* ($\mu = 1.5$, $\sigma = 1.6$), and *OkaySingleUse* ($\mu = 1.8$, $\sigma = 1.5$). They generally showed agreement even on the statement about any person using the technology: *OkayAnyUse* ($\mu = 1.0$, $\sigma = 1.6$). We also noticed that our online participants tended to disagree on the negative statements: *PrivacyConcerns* ($\mu = -0.4$, $\sigma = 1.8$), *UncomfortableAssistUse* ($\mu = -1.4$, $\sigma = 1.7$), *UncomfortableRecording* ($\mu = -0.1$, $\sigma = 2.1$), and *UncomfortableSingleUse* ($\mu = -1.3$, $\sigma = 1.9$).

Figure 8.4 visualizes their responses to the scenario questions. We observed that the less features the assistive camera system detected from pedestrians, the more positive participants



(a) Participants in Glasses felt more uncomfortable about camera recording than participants in GoPro.

(b) Participants in Glasses felt more uncomfortable about single data use than participants in GoPro.

(c) Participants became more negative about any people using it after watching the UserVideo.

Figure 8.5: Comparisons of online participants' responses.

tended to be about the system. Furthermore, more than 85% of the participants answered that they would be okay with the system recognizing them if they knew the users. As for the question about *Recognition of any people*, more than 30% of the participants were amicable to the system recognizing people even if they did not know users, but more than half the participants disagreed to this.

8.3.1.2 Camera visibility

Comparing responses from the Glasses and GoPro, we then sought to see whether camera visibility affects people's perceptions of the technology. We first counted the number of participants who agreed that it was easy to notice a camera in the passerby video; this question was asked right after they watched the passerby perspective video. Of the 90 participants in GoPro, 78 (87%) agreed that it was easy to notice the camera while only 14 (16%) of the 90 participants in Glasses agreed on it. Using the Mann-Whitney-Wilcoxon test for significance testing and the Rosenthal's formula for effect size calculation, we found that the difference in camera visibility

between the cameras was high and significant ($p < 0.01$, $r = 0.715$).

We further analyzed their responses to the social acceptance questions asked after the passerby video to see effects of camera visibility on their perceptions of the technology. Figure 8.5(a) and 8.5(b) demonstrate that participants in Glasses agreed to feeling more uncomfortable about a wearable camera that either records them ($p < 0.05$, $r = 0.176$) or uses images/videos for detection without saving ($p < 0.05$, $r = 0.152$) than participants in GoPro were.

8.3.1.3 Awareness of attributes detected by technology

Social acceptance of wearable camera technology has been primarily evaluated from a bystander's view point, meaning that potential bystanders may not know about what data the technology will take to accomplish its objectives. To evaluate effects of awareness of the data processed by the wearable technology on social acceptance, we examined participants' responses to the social acceptance questions repetitively asked before and after the user perspective video. Note that the user video informed of the wearable camera's function of capturing visual attributes of pedestrians. For this analysis, all the responses from the Glasses and GoPro groups were merged and used, and the Wilcoxon signed-rank test was used to compare their responses before and after the user video.

As illustrated in Figure 8.6, participants generally tended to be positive toward wearable camera technology if it was designed for assistive purpose after watching the passerby video and the user video. However, we observed that more participants tended to become uncomfortable seeing people using such a wearable camera device in public spaces, even if it was for assistive purpose, after the user video (*UncomfortableAssistUse*). The number of participants who agreed

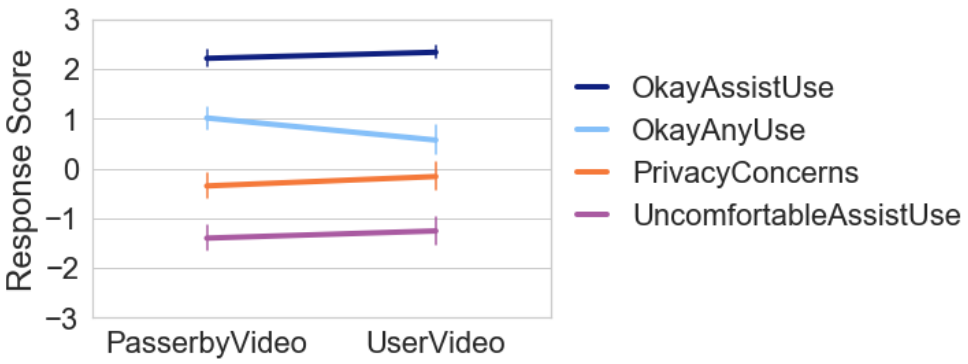


Figure 8.6: Online sighted participants' responses to the social acceptance questions before and after the user perspective video. Responses are converted to numeric values (-3 to 3): -3=*Strongly disagree*, 0=*Neutral*, 3=*Strongly agree*.

to feeling uncomfortable increased after the user video (from 38 to 50 out of the 180 participants), but the difference was not statistically significant. For one category, *OkayAnyUse*, participants became more negative toward it after they watched the user video ($p < 0.01$, $r = 0.265$); Figure 8.5(c) shows their responses before and after the user video. In particular, one online participant strongly agreed that any people can use such technology by saying "people can wear whatever they want" before the user video. However, after learning about visual features estimated by the technology, the participant switched to strongly disagreeing with this category and commented "I would feel like they're invading my privacy."

Lastly, we questioned whether camera visibility affects the participants' perceptions of the technology before and after they are aware of data processed by the technology. Participants in Glasses and GoPro agreed to change their perceptions of the technology after watching the user perspective video, but participants in Glasses ($\mu = 1.41$, $\sigma = 1.54$) agreed more to change their perceptions than did participants in GoPro ($\mu = 0.73$, $\sigma = 1.87$): $p < 0.01$, $r = 0.182$. Using ANCOVA, we checked if there was a significant difference in perception changes due to camera visibility. The result indicates that there is a strong interaction between the camera visibility and

the perception changes; *i.e.*, the camera visibility affected the changes in their opinions on social acceptance of wearable technology before and after the user video ($p < 0.05$).

8.3.2 In-Situ Study

8.3.2.1 System performance measured with log data

We first measured the performance of visual feature estimations of our prototype system to check if feedback from the system properly informed blind participants of the pedestrians' attributes. For this measurement, we manually annotated images that our system used to generate and deliver feedback messages about pedestrians to blind participants. The system achieved, on average, f-score of 0.772, 0.778, 0.775, and 1.0, on estimations of name, gender, head pose, and position, respectively. The position estimation achieved the maximum performance, meaning that, whenever the prototype system detected a face, the detected face was a pedestrian's and thus estimated the face position, correctly.

8.3.2.2 Perspectives of blind users

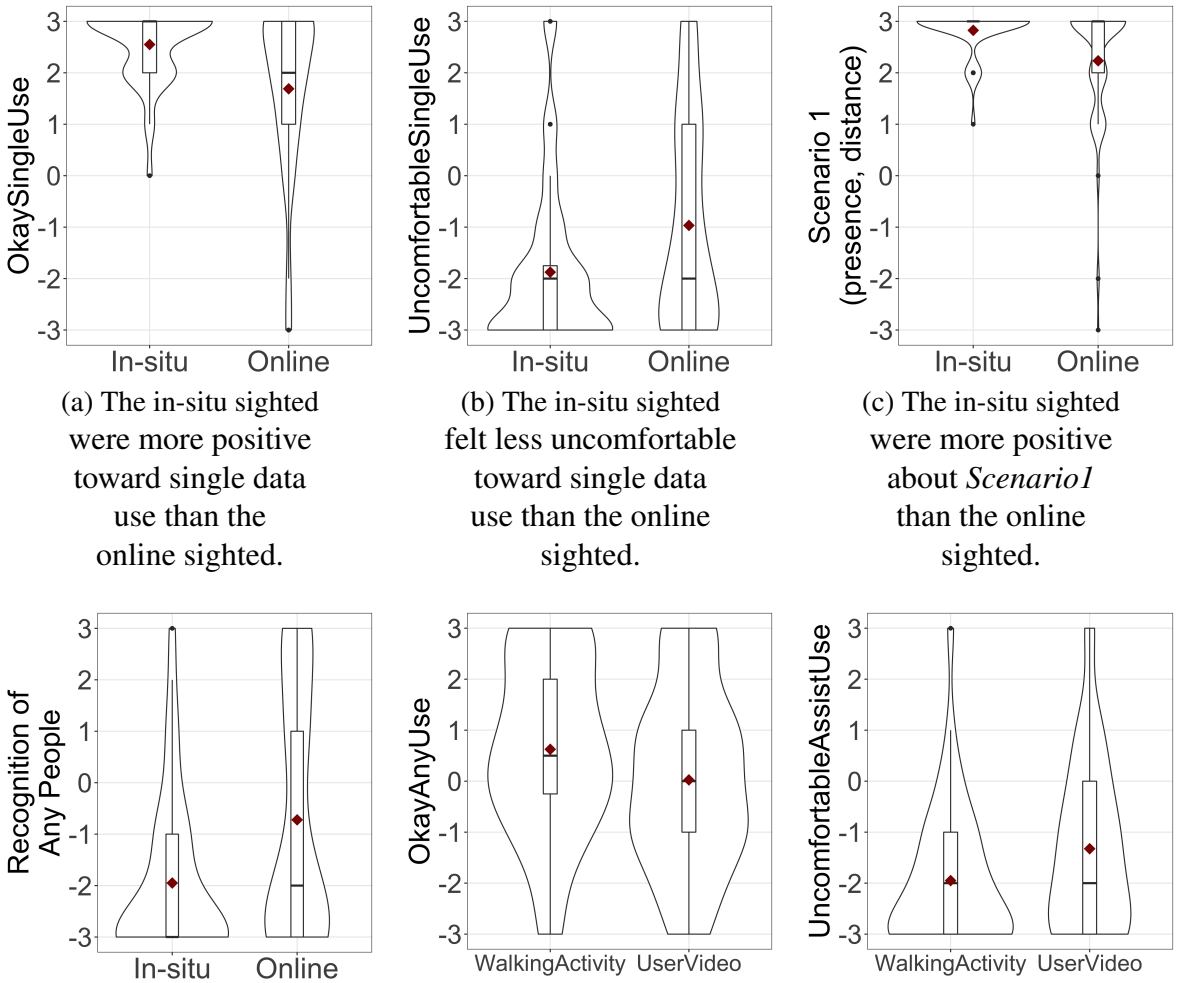
In the post-study session, we observed that some blind participants only thought that the gender estimation was somewhat problematic; although the f-scores of the name and head pose estimations are similar to the gender estimation's. It seems that this discrepancy may have been due to different levels of attention to visual features, since we learned that some blind participants focused on gender estimation to appropriately address unknown pedestrians using either *sir* or *ma'am* when initiating a conversation; *e.g.*, P6 said, "(when the system detected a man) I said 'excuse me, sir' and it's a woman, ... that's sort of embarrassing."

Furthermore, we noticed that blind participants tended to be moderately positive about assistive wearable cameras according to their responses to the questions described in Table 8.3. They tended to be positive about using such systems in public (*OkayUseInPublic*) and agreed that it could help them interact with not only known people but also unknown people (*HelpInteractWithKnown*, *HelpInteractWithUnknown*). Blind participants ($\mu = -1.0$, $\sigma = 2.0$) tended to disagree more with there being privacy concerns about the assistive technology than sighted participants (online (Glasses): $\mu = -0.1$, $\sigma = 1.8$, in-situ: $\mu = 0.2$, $\sigma = 1.8$). Also, they generally agreed to willingly provide more photos/videos of known people to improve recognition system, so long as those known people agree.

8.3.2.3 Perspectives of sighted participants

We attempted, both in online and in-situ studies, to have sighted people who are interacting often with technology either as Turkers (online) or as students, faculty, and visitors in a technological institution (in-situ), and thus include participants who tend to be positive toward technology in general. Our rationale behind this decision, beyond the obvious practicality, was to reach out to an audience seen as potential early adopters of technology. Using responses from in-situ and online sighted participants, we further evaluated social acceptance of an assistive wearable camera from perspectives of sighted passerby. Since the in-situ participants personally experienced the smartglasses in the user study, responses from the online participants who remotely experienced the smartglasses ($n = 90$) were used to control variances that can be caused by the camera form.

After the walking activity in the in-situ study, except for *UncomfortableRecording* ($\mu = 1.0$, $\sigma = 1.5$), the in-situ participants seemed to be generally positive about the assistive technol-



(a) The in-situ sighted were more positive toward single data use than the online sighted.

(b) The in-situ sighted felt less uncomfortable toward single data use than the online sighted.

(c) The in-situ sighted were more positive about *Scenario1* than the online sighted.

(d) The in-situ sighted were more negative about any people recognition than the online sighted.

(e) The in-situ sighted became more negative about any people using it after the user video.

(f) The in-situ sighted felt more uncomfortable about the assistive camera after watching the user video.

Figure 8.7: The four figures on the left for comparison of responses from in-situ sighted participants with those from online sighted participants, and two figures on the right for comparison of in-situ sighted participants' responses before and after the user video.

ogy in the other seven categories: *OkayAssistUse* ($\mu = 2.4$, $\sigma = 1.0$), *OkayAnyUse* ($\mu = 0.6$, $\sigma = 1.7$), *PrivacyConcerns* ($\mu = -0.1$, $\sigma = 1.8$), *UncomfortableAssistUse* ($\mu = -2.0$, $\sigma = 1.4$), *OkayRecording* ($\mu = 1.0$, $\sigma = 1.8$), *OkaySingleUse* ($\mu = 2.6$, $\sigma = 0.7$), and *UncomfortableSingleUse* ($\mu = -1.9$, $\sigma = 1.6$). As for *UncomfortableRecording*, both the in-situ and online (Glasses) participants tended to feel uncomfortable about camera recording (online (Glasses):

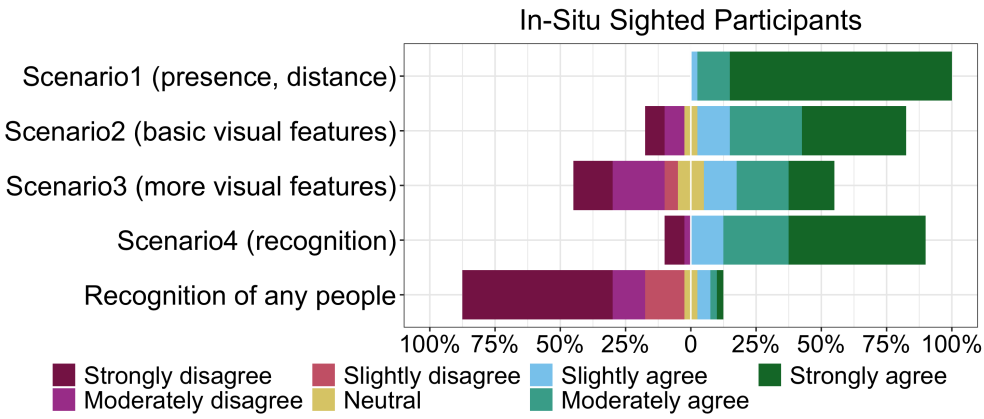


Figure 8.8: In-situ sighted participants' opinions on use cases of assistive wearable cameras. Participants were asked whether they would be okay with each of the use cases.

$\mu = 0.3$, $\sigma = 2.0$), and there was no significant difference between these two groups. On the other hand, we found that the in-situ participants had more positive attitudes toward the single use of images/videos in the assistive technology than did the online participants. As depicted in Figure 8.7(a) and 8.7(b), the in-situ participants were more positive about the assistive technology using images/videos for one-time detection ($p < 0.01$, $r = 0.272$) and felt less uncomfortable about it ($p < 0.05$, $r = 0.206$) than the online.

Moreover, Figure 8.8 visualizes the in-situ participants' responses to the questions about the scenarios introduced in the user video. We compared these responses with the ones from the online participants (Glasses) to see if there was any significant difference between the groups. As described in Figure 8.7(c), we found that the in-situ participants were generally more positive toward the technology detecting their presence and distance between blind users and them, *Scenario1 (presence, distance)*, than the online participants were ($p < 0.01$, $r = 0.273$). In addition, although both of the groups were negative toward the system recognizing anyone (as opposed to just known people), Figure 8.7(d) show that the in-situ participants were more negative about being recognized when they do not know blind users ($p < 0.01$, $r = 0.246$).

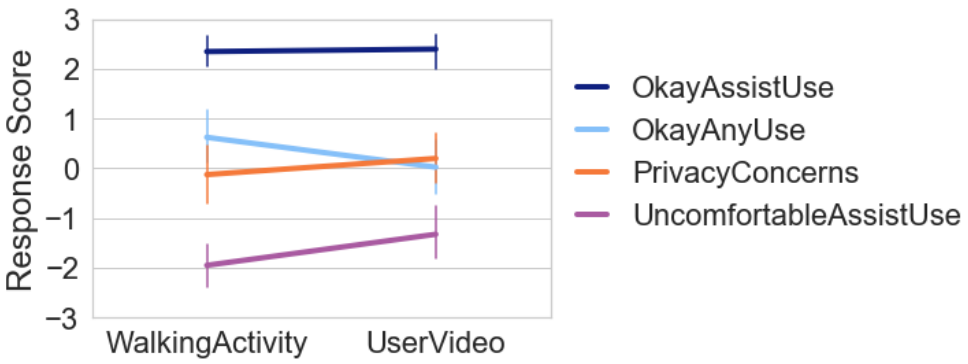
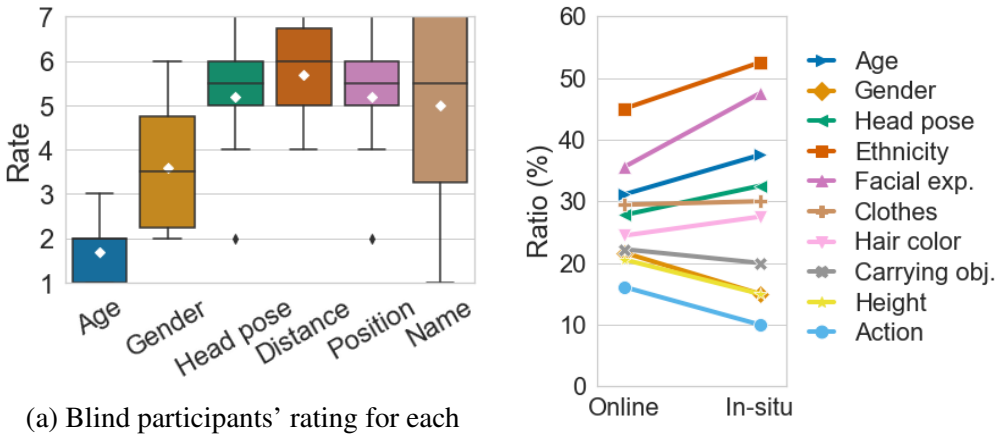


Figure 8.9: In-situ sighted participants' responses to the social acceptance questions before and after the user video. Responses are converted to numeric values (-3 to 3): -3=*Strongly disagree*, 0=*Neutral*, 3=*Strongly agree*.

8.3.2.4 Awareness of attributes detected by technology

Analyzing in-situ sighted participants' responses to the social acceptance questions before and after the user video, we also evaluated effects of awareness of the data used in the technology on the in-situ sighted participants' perceptions of the technology; they learned about what such an assistive system may detect from pedestrians by watching the user video in the post-study session. Figure 8.9 shows their responses to the questions before and after the user video. Similar to our finding from the online survey data, we observed that in-situ sighted participants grew more negative toward anyone using a wearable camera system (*OkayAnyUse*) after the user video ($p < 0.05$, $r = 0.282$) as shown in Figures 8.7(e) and 8.9. Before the user video, one in-situ participant mentioned, "it doesn't cause any inconvenience to others," but, after the user video, that same participant commented, "they may use it to analyze our personal information for some bad purpose"; this participant changed the response from *Strongly agree* to *Moderately disagree*.

As shown in Figure 8.7(f), we found that the in-situ participants felt more uncomfortable about blind people using the assistive camera (*UncomfortableAssistUse*) after watching the user



(a) Blind participants' rating for each feature detection. The top-3 preferred features are distance, head pose, and position. (1=the least preferred, 7=the most preferred)

(b) Ratio of sighted participants for each feature that do not like to be detected by such assistive system.

Figure 8.10: Participants' opinions on feature detection.

video and realizing that their visual features were detected by the system ($p < 0.05$, $r = 0.253$), which was not observed from the online participants. One in-situ participant said "it's okay for me" and strongly disagreed on *UncomfortableAssistUse* before watching the user video, but, after the video, mentioned "when I realize it will collect too much information, I will feel a little bit uncomfortable" and changed to slightly agreeing to feel uncomfortable.

8.3.2.5 Blind users and sighted pedestrians

Lastly, we further investigated conflicts of interest in functionalities of an assistive wearable camera from the perspectives of blind users and sighted passerby. Blind participants were asked to rate six visual features (age, gender, head pose, distance, position, name) that the prototype system provided them with during the user study. The rating scale was 1 to 7 where 1 indicated the least important feature and 7 indicated the most important feature; redundancy was allowed for the rating. On the other hand, we asked sighted participants which of their visual features

they would prefer not to be detected and reported by such an assistive system.

Figure 8.10(a) shows that blind participants preferred to obtain information about *head pose*, *distance*, and *position* of pedestrians, while *age* is the last feature that blind participants did care about. P4 mentioned, "if there is a baby one year old, (the baby) can't tell you what the number on a door is. Other than that, (age estimation) doesn't make any difference." Figure 8.10(b) shows the ratio of sighted participants for each feature that they would not want be exposed by the assistive system. *Ethnicity*, *facial expression*, and *age* are the top-3 features that participants would not want such a system to detect from them. Yet, some sighted participants (31% of online sighted participants and 23% of in-situ sighted participants) answered that they would allow an assistive system to detect any necessary visual features from themselves.

Features that blind participants did not care to obtain from such an assistive system mostly match those that sighted participants did not want to provide via the system; *age* and *ethnicity* are the features that sighted participants did not want the system to detect, and blind participants did not think that the system needed to provide. On the other hand, most of the blind participants considered *head pose* as an essential feature that such a system needed to provide because they can use this information as an indicator for pedestrian's availability for social interactions. In particular, P10 remarked, "now people with cell phones, they are talking on their phones, but (blind people) don't know whether they are talking to you or on a phone. The key is if they are looking at (blind people)." However, around 30% of each group of sighted participants did not want to provide this information to the system. Also, it seems that *facial expression* could be an additional indicator for blind users to decide whether to begin social interactions with pedestrians; P5 said, "it might be nice to have facial expressions. If someone is upset, I might not want to ask them for assistance." However, *facial expressions* was the second feature that sighted participants

(36% of the in-situ and 48% of the online) did not want the system to detect from them.

8.4 Discussion

8.4.1 Implications

We disclose that there were conflicting features that blind users would need from the assistive technology but sighted pedestrians might not want be detected by assistive wearable cameras. These include head pose and facial expression. In addition, some of the blind participants commented that they do not want such a system to be an indication of their disability; P7 mentioned, "I wouldn't wear heavier, bulkier smartglasses because it's more likely to draw unwanted attention." On the other hand, our analysis with sighted participants indicates that camera visibility and awareness of the technology may affect pedestrians' perceptions of such an assistive wearable camera; the less visible a camera was, the more negative the response of sighted participants toward the technology became after discovering its image/video use. One sighted participant said, "I'm not looking to be recorded without my knowledge, so it's important that it is discernible that the blind individual is wearing a camera.". It seems as if the more visible camera might have caused people to prepare for being recorded. The less visible camera (smartglasses), on the other hand, might have led sighted people to unpleasantly learn about its photo-taking or video recording function after they already had been recorded. Hence, conflicts of interest in such technology should first be collected, carefully reviewed, and appropriately addressed for all stakeholders.

Our analysis on responses from online and in-situ sighted participants provides further evidence on the prior findings that people are willing to accept wearable camera technology for assistive purposes [16]. However, the differences in the online and in-situ sighted participants'

responses to *Recognition of any people* indicate that in-person experience may engender different effects on people's perception of such technology than remote experiences do. It indeed reconfirms the prior finding that in-situ (direct) experience catalyzes people's perception of technology more than remote (indirect) experience did in the user study [251]. Although we acknowledge the high cost of conducting on-site user study, we encourage researchers to combine on-site user studies with online user studies to better understand people's opinion on technology.

8.4.2 Limitations

There are, however, several limitations to our work. First, our in-situ participants might not represent the true population due to their skewed demographic information, such as age and educational background. Among both blind and sighted participants in the in-situ user study, the age distribution was skewed (blind participants: 49–73 years, and sighted participants: 19–39 years). We see similar trends in their attitudes toward assistive wearable camera technology even when controlling for educational background and comparing the 40 in-situ sighted participants to 53 online sighted participants, who watched the smartglasses videos and indicated having a bachelor's or advanced degree, but it may not indicate that there is no demographic or confounding factor that may have led to the differences in social acceptance between the two groups of sighted participants. We believe that conducting a user study with more participants with various backgrounds might address such issues and help us understand social acceptance, more generally.

Moreover, the scenario of detecting pedestrians in a corridor was chosen to be the main use case for our working prototype system, since it allowed us to conduct user study and measure social acceptance of such technology in a controlled environment. This single scenario enables

repeatability of our user study, but could limit exploration of other factors in social acceptance due to its lack in representability of a real-world use case. As such technology could reveal different social tensions in different scenarios, our future work will be to explore further use cases of assistive wearable cameras to evaluate its social acceptance in general.

8.5 Summary

Rise of wearable camera technology has triggered researchers' attention to its social acceptance and potential assistive uses. Considering not only users and bystanders but also several other factors that may affect people's perceptions of such technology, our work explored social acceptance of a potential use, *pedestrian detection*, of a wearable camera system for blind people. We found that there was agreement and disagreement between blind users and sighted pedestrians on what such a system should and should not provide, and observed that both the camera visibility and the way of experiencing the technology may affect bystanders' attitudes toward the technology. We also reconfirmed that people were more positive toward wearable cameras if it was designed for assistive purposes. Our findings suggest that it would be necessary to evaluate social acceptance of such technology in a more realistic environment and understand the needs of not only users but also people who would be involved in the technology.

Chapter 9: Conclusion

This dissertation explores the potential and challenges of using AI for helping blind users access to their visual surroundings. Using two different camera form factors, smartphone camera and smartglasses camera, we contribute to (i) exploring the data scarcity problem in employing AI for assistive technologies, (ii) designing AI based on observations from target users' interactions with it, (iii) understanding error-prone AI in the context of machine teaching, and (iv) studying social acceptance and privacy concerns that may arise due with assistive cameras.

The inclusion of the intended user group, blind users in this work, is found to be essential for collecting training data; such data are critical for employing AI-infused systems that work in the real-world and that respond to blind users' intentions. However, the nature of AI making errors needs to be considered in the design of interactive AI systems. Furthermore, different camera form factors may affect blind users' behaviors, such as camera aiming, and lead to different societal issues regarding assistive camera technologies. We believe that it is necessary to study such societal issues prior to a system design so as to incorporate findings in the actual system.

9.1 Contributions

This section discusses contributions with respect to both understanding the potential and challenges of blind people's interactions with assistive computer vision and exploring different

camera form factors and societal issues that such assistive camera may engender. This thesis has (i) identified the problems of employing AI that is trained by other population than the target population, (ii) studied blind users' interaction with camera devices for designing AI that incorporates those interactions, (iii) revisited prior research work on photography of blind people using AI on two different camera form factors (smartphones and smartglasses), and (iv) explored blind users' interactions with those different camera form factors and societal issues that may be led by such form factors.

The earlier chapters have shown that it is essential to understand blind users' interactions with AI and a camera form factor to devise assistive camera technologies from which blind users can benefit. Their interactions with smartphone cameras have been widely studied and also revisited in our work, but little work focuses on blind users' interactions with smartglasses in the context of object recognition. Our proposed work therefore tries to learn their interactions with the wearable camera form and find insightful observations that can lead to design implications for future assistive systems. Deploying our prototype system on smartglasses in the wild, we expect to (i) learn blind users' interactions with teachable smartglasses and (ii) evaluate the smartglasses for object recognition performance and social acceptance.

Data collection during the deployment study would allow us to identify patterns in their camera aiming and compare those patterns with the ones with the smartphone camera. Focusing on their camera interactions in the context of object recognition, we believe that data would reveal how blind users ensure their objects are captured by their smartglasses. Also, although an AI-powered teachable interface enables blind users to teach their objects to the smartglasses, AI is an error-prone system by nature. The validation phase in the training procedure is added to help blind users perceive potential errors in their personalized object recognizer and address

them by iterating the training procedure with data that leads to the errors. Our proposed study thus can expand our understanding of blind people's photography using the wearable camera and perception of machine teaching.

As blind users are wearing the camera as a form of glasses, the camera is naturally affected by their head pose. This nature may change their camera aiming compared to their camera aiming with smartphones, and thus result in different visual characteristics in their captured data. As our proposed study deploys our prototype system to collect data with blind users in real-world situations, it would gauge the capability of AI in real-world applications and uncover challenges that are not found in in-lab or controlled data collection. Moreover, the always-on feature of the camera on smartglasses may raise privacy concerns as it can capture unwanted or private visual information and other people, accidentally. During the deployment study, we plan on having semi-structured interviews with blind participants and their house members. Blind participants' responses would allow us to understand their perception and concerns regarding teachable smartglasses and explore its potentials, such as potential applications to which such assistive smartglasses are best suited. On the other hand, feedback from their house members would help us investigate societal issues raised by the use of such assistive wearable cameras in the real world. We believe that it is vital to understand social perceptions of such assistive technology as the society may not accept assistive technology that invades other people's privacy even if it can enhance the accessibility for blind people.

9.2 Future Directions

As in many other applications, artificial intelligence (AI) has been considered as an additional module in assistive technologies. We envision that more users will be interacting with assistive systems that employ AI. Our exploratory work on employing AI in assistive technologies for blind people, which can help them access their visual surroundings, suggests several opportunities for future work on assistive AI research. We highlight and discuss design implications that future research may consider.

9.2.1 Considering Different Stakeholders in Various Assistive Applications

This dissertation has emphasized the importance of understanding societal issues that assistive camera systems may engender as it can affect the design of such assistive technologies directly. As our initial investigation of social acceptance (Chapter 8) suggests that different stake holders may have different opinions on access technologies, considering their perspectives plays an important role in designing access technologies. Anecdotal observation from our user study with smartglasses for teachable object recognition (Chapter 6) reveals that blind users may not want to keep data, which they capture in their home environments, on cloud as it can be susceptible to data breaches. Simultaneously, their household members may not accept an assistive camera if it captures them (*e.g.*, their face) unknowingly and stores the data on a place that others can access with or without appropriate permission. Future work should look to explore various factors that induce such societal issues (*e.g.*, privacy concerns) to expand our knowledge and understanding of what should be considered in its design or not. Different factors may be involved in different scenarios and applications and thus need to be carefully examined prior to the design

and deployment of access technology.

9.2.2 Exploring User Interactions with Teachable Interfaces in the Wild

Throughout our user studies using smartphones (Chapter 4) and smartglasses (Chapter 6), we explore teachable interfaces that blind people use to access the visual information of their objects. Although our smartglasses study simulated a deployment study, understanding the end-user's interaction with teachable interfaces in the real world has been still unexplored. In addition, teachable interfaces have been studied with different user groups (*e.g.*, people with dysarthric speech and physical disabilities [228], children [263], and people who are deaf or hard of hearing [264]). Different user groups may want to use such teachable interfaces in different scenarios, which thus need to be adequately studied to understand their demands with respect to teachable interfaces. Investigating teachable interfaces in a real-world environment can enable researchers and designers to observe user interactions that can suggest more practical design implications. For example, deploying such a teachable camera system could help researchers study further on how users interact with such a teachable system, in what situation they would like to use, and what challenges still remain to be addressed to make such a teachable interface more accessible to end-users. Indeed, teachable interfaces have been realized in commercial assistive products designed for blind users, such as OrCam [10], Envision [47], and SeeingAI [44]. Given that teachable interfaces have been actively researched in academia [4, 57, 160, 189] and real-world data have been collected by industry, we see that future research can greatly benefit from the collaboration and synergy between academia and industry.

9.2.3 Making Evaluation of Computer Vision Accessible

A potential issue with teachable interfaces is that what a teachable interface learns from users may not be transparent and accessible to them. We have observed it during our smartglasses study, especially when participants spotted wrong recognition outputs from their personalized object recognizer. Participants wanted to know what caused this failure and why their recognizer thought that the object was something else, not the one they were expecting. When teaching something to an AI-infused system, users may want to examine their system, which resembles a teacher giving an exam to students in school. The ability of auditing AI has the potential to help the user understand the capabilities of their teachable systems and improve their teaching strategies accordingly. As of now, many researchers have worked on making what AI models learn transparent, but little work has studied on how to connect those fundamental research work to human-AI interactions. It highlights great research directions that human-AI researchers can take for their future work.

9.3 Epilogue

This dissertation has explored assistive teachable interfaces for blind people considering various factors such as camera form factors, applications, user roles, and societal issues. Artificial intelligence holds promise in making assistive systems personalized to blind user and promoting them as an active directory of such technologies. Our exploratory work suggests that understanding blind users' interactions with teachable interfaces needs to be prioritized in the design of such AI-infused technologies. We see that different stakeholders may have different perspectives towards assistive AI-infused technologies and thus advocate further research on studying various

factors that can be associated with data, privacy, and social acceptance. This dissertation serves as exploratory work on these emerging research topics, suggesting various opportunities for future research.

Bibliography

[1] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *arXiv preprint arXiv:1802.08218*, 2018.

[2] Kyungjun Lee and Hernisa Kacorri. Hands holding clues for object recognition in teachable machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[3] Hernisa Kacorri. Teachable machines for accessibility. *SIGACCESS Access. Comput.*, (119):10–18, November 2017.

[4] Hernisa Kacorri, Kris M Kitani, Jeffrey P Bigham, and Chieko Asakawa. People with visual impairment training personal object recognizers: Feasibility and challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5839–5849. ACM, 2017.

[5] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2117–2126. ACM, 2013.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[8] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. Crowdsourcing the perception of machine teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[9] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE, 2018.

[10] OrCam MyEye 2. Help people who are blind or partially sighted, 2020.

[11] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 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*, page 215. ACM, 2018.

[12] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–42, 2015.

[13] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–60, 2017.

[14] Shi Qiu, Jun Hu, and Matthias Rauterberg. Nonverbal signals for face-to-face communication between the blind and the sighted. In *Proceedings of International Conference on Enabling Access for Persons with Visual Impairment*, pages 157–165, 2015.

[15] Halley P Profita, James Clawson, Scott Gilliland, Clint Zeagler, Thad Starner, Jim Budd, and Ellen Yi-Luen Do. Don't mind me touching my wrist: a case study of interacting with on-body technology in public. In *Proceedings of the 2013 International Symposium on Wearable Computers*, pages 89–96. ACM, 2013.

[16] Halley Profita, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun K Kane. The at effect: how disability affects the perceived social acceptability of head-mounted display use. In *proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4884–4895. ACM, 2016.

[17] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P Bigham. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 203–210. ACM, 2011.

[18] Michael Brock and Per Ola Kristensson. Supporting blind navigation using depth sensing and sonification. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 255–258. ACM, 2013.

[19] Jan Balata, Zdenek Mikovec, and Lukas Neoproud. Blindcamera: Central and golden-ratio composition for blind photographers. In *Proceedings of the Mulitimedia, Interaction, Design and Innnovation*, MIDI '15, New York, NY, USA, 2015. Association for Computing Machinery.

[20] Jeffrey P Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. Vizwiz:: Locateit-enabling blind people to locate objects in their environment. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 65–72. IEEE, 2010.

[21] Alexander Fiannaca, Ilias Apostolopoulous, and Eelke Folmer. Headlock: A wearable navigation aid that helps blind cane users traverse large open spaces. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers and Accessibility*,

ASSETS '14, page 19–26, New York, NY, USA, 2014. Association for Computing Machinery.

[22] Annuska Zolyomi, Anushree Shukla, and Jaime Snyder. Technology-mediated sight: A case study of early adopters of a low vision assistive technology. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 220–229. ACM, 2017.

[23] Lee Stearns, Leah Findlater, and Jon E Froehlich. Design of an augmented reality magnification aid for low vision users. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 28–39. ACM, 2018.

[24] Aira. Your life, your schedule, right now., 2017.

[25] Bjørn Karmann. Objectifier spatial programming, 2017.

[26] Julie Rico and Stephen Brewster. Gestures all around us: user differences in social acceptability perceptions of gesture based interfaces. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*, page 64. ACM, 2009.

[27] Calkin S Montero, Jason Alexander, Mark T Marshall, and Sriram Subramanian. Would you do that?: understanding social acceptance of gestural interfaces. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, pages 275–278. ACM, 2010.

[28] Taslima Akter, Tousif Ahmed, Apu Kapadia, and Swami Manohar Swaminathan. Privacy considerations of the visually impaired with camera based assistive technologies: Misrepresentation, impropriety, and fairness. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14, 2020.

[29] Marynel Vázquez and Aaron Steinfeld. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 95–102. ACM, 2012.

[30] Yu Zhong, Pierre J Garrigues, and Jeffrey P Bigham. Real time object scanning using a mobile phone and cloud-based visual search engine. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, page 20. ACM, 2013.

[31] Kyungjun Lee, Jonggi Hong, Simone Pimento, Ebrima Jarjue, and Hernisa Kacorri. Revisiting blind photography in the context of teachable object recognizers. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 83–95, 2019.

[32] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342. ACM, 2010.

[33] Digit-Eyes. Identify and organize your world., 2010.

[34] EyeNote. Mobile device application to denominate federal reserve notes (u.s. paper currency) as an aid for the blind or visually impaired to increase accessibility, 2010.

[35] LookTel Recognizer. Instantly recognize everyday objects., 2012.

[36] TapTapSee. Mobile camera application designed specifically for the blind and visually impaired ios users., 2012.

[37] Talking Goggles. A camera with speech., 2013.

[38] CamFind. Search the physical world., 2013.

[39] EyeSpy. The world's best object recognition mobile app, 2015.

[40] BeMyEyes. Lend you eyes to the blind., 2015.

[41] BeSpecular. Let blind people see through your eyes., 2016.

[42] Aipoly Vision. Sight for blind & visually impaired., 2016.

[43] Joan Sosa-García and Francesca Odone. "hands on" visual recognition for visually impaired users. *ACM Transactions on Accessible Computing (TACCESS)*, 10(3):8, 2017.

[44] Seeing AI. A free app that narrates the world around you., 2017.

[45] VocalEyes AI. Computer vision for the blind., 2017.

[46] WayAround. The smart assistant for people who are blind., 2018.

[47] Envision AI. Enabling vision for the blind., 2018.

[48] Roberto Manduchi and James Coughlan. (computer) vision without sight. *Communications of the ACM*, 55(1):96–104, 2012.

[49] Rabia Jafri, Syed Abid Ali, Hamid R Arabnia, and Shameem Fatima. Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. *The Visual Computer*, 30(11):1197–1222, 2014.

[50] Marco Leo, G Medioni, M Trivedi, Takeo Kanade, and Giovanni Maria Farinella. Computer vision for assistive technologies. *Computer Vision and Image Understanding*, 154:1–15, 2017.

[51] Jeremi Sudol. *LookTel—Computer Vision Applications for the Visually Impaired*. PhD thesis, UCLA, 2013.

[52] i.d. mate. Talking bar code scanners., 2018.

[53] Opticon. Handheld scanner., 2018.

[54] KNFB Reader. Access to print materials., 2018.

[55] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1153–1162, 2016.

[56] NantMobile Money Reader. Instantly recognizes currency and speaks the denomination, 2017.

[57] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. Orbit: A real-world few-shot dataset for teachable object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10818–10828, 2021.

[58] Yongsik Jin, Jonghong Kim, Bumhwi Kim, Rammohan Mallipeddi, and Minho Lee. Smart cane: Face recognition system for blind. In *Proceedings of the 3rd International Conference on Human-Agent Interaction*, pages 145–148. ACM, 2015.

[59] Priyan Malarvizhi Kumar, Ushadevi Gandhi, R Varatharajan, Gunasekaran Manogaran, R Jidhesh, and Thanjai Vadivel. Intelligent face recognition and navigation system using neural learning for smart security in internet of things. *Cluster Computing*, pages 1–12, 2017.

[60] M Saquib Sarfraz, Angela Constantinescu, Melanie Zuzej, and Rainer Stiefelhagen. A multimodal assistive system for helping visually impaired in social interactions. *Informatik-Spektrum*, 40(6):540–545, 2017.

[61] Lee Stearns and Anja Thieme. Automated person detection in dynamic scenes to assist people with vision impairments: An initial investigation. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '18, page 391–394, New York, NY, USA, 2018. Association for Computing Machinery.

[62] SeeingAI. Talking camera app for those with a visual impairment, 2020.

[63] Katrin Wolf, Yomna Abdelrahman, David Schmid, Tilman Dingler, and Albrecht Schmidt. Effects of camera position and media type on lifelogging images. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, pages 234–244, 2015.

[64] David Guth and Robert LaDuke. The veering tendency of blind pedestrians: An analysis of the problem and literature review. *Journal of Visual Impairment & Blindness*, 1994.

[65] Eli Peli and Russell L Woods. Image enhancement for impaired vision: the challenge of evaluation. *International Journal on Artificial Intelligence Tools*, 18(03):415–438, 2009.

[66] Yuhang Zhao, Michele Hu, Shafeka Hashash, and Shiri Azenkot. Understanding low vision people's visual perception on commercial augmented reality glasses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4170–4181, 2017.

[67] Lee Stearns, Victor DeSouza, Jessica Yin, Leah Findlater, and Jon E Froehlich. Augmented reality magnification for low vision users with the microsoft hololens and a finger-worn camera. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 361–362, 2017.

[68] Erin L Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P Bigham. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1225–1236. ACM, 2013.

[69] Dustin Adams, Sri Kurniawan, Cynthia Herrera, Veronica Kang, and Natalie Friedman. Blind photographers and vizsnap: A long-term study. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 201–208. ACM, 2016.

[70] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194, 2001.

[71] Susumu Harada, Daisuke Sato, Dustin W Adams, Sri Kurniawan, Hironobu Takagi, and Chieko Asakawa. Accessible photo album: enhancing the photo sharing experience for people with visual impairment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2127–2136. ACM, 2013.

[72] Dustin Adams, Lourdes Morales, and Sri Kurniawan. A qualitative study to support a blind photography mobile application. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, page 25. ACM, 2013.

[73] Nadia Gosselin-Kessiby, John F Kalaska, and Julie Messier. Evidence for a proprioception-based rapid on-line error correction mechanism for hand orientation during reaching movements in blind subjects. *Journal of Neuroscience*, 29(11):3485–3496, 2009.

[74] Anhong Guo, Xiang'Anthony' Chen, Haoran Qi, Samuel White, Suman Ghosh, Chieko Asakawa, and Jeffrey P Bigham. Vizlens: A robust and interactive screen reader for interfaces in the real world. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 651–664. ACM, 2016.

[75] Sergio Mascetti, Andrea Gerino, Cristian Bernareggi, Silvia D'Acquisto, Mattia Ducci, and James M Coughlan. Justpoint: Identifying colors with a natural user interface. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 329–330. ACM, 2017.

[76] Róisín McNaney, John Vines, Daniel Roggen, Madeline Balaam, Pengfei Zhang, Ivan Poliakov, and Patrick Olivier. Exploring the acceptability of google glass as an everyday assistive device for people with parkinson's. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 2551–2554. ACM, 2014.

[77] Qingguo Xu, Sen-ching Samson Cheung, and Neelkamal Soares. Littlehelper: an augmented reality glass application to assist individuals with autism in job interview. In *2015*

Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pages 1276–1279. IEEE, 2015.

[78] Meethu Malu and Leah Findlater. Personalized, wearable control of a head-mounted display for users with upper body motor impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 221–230. ACM, 2015.

[79] Shurug Al-Khalifa and Muna Al-Razgan. Ebsar: Indoor guidance for the visually impaired. *Computers & Electrical Engineering*, 54:26–39, 2016.

[80] Philipp A Rauschnabel and Young K Ro. Augmented reality smart glasses: An investigation of technology acceptance drivers. *International Journal of Technology Marketing*, 11(2):123–148, 2016.

[81] eSight. Electronic glasses for the legally blind, 2020.

[82] Tousif Ahmed, Apu Kapadia, Venkatesh Potluri, and Manohar Swaminathan. Up to a limit? privacy concerns of bystanders and their willingness to share additional information with visually impaired users of assistive technologies. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–27, 2018.

[83] Kristen Shinohara and Jacob O Wobbrock. In the shadow of misperception: assistive technology use and social interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 705–714. ACM, 2011.

[84] Hanlu Ye, Meethu Malu, Uran Oh, and Leah Findlater. Current and future mobile and wearable device use by people with visual impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3123–3132. ACM, 2014.

[85] Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3523–3532. ACM, 2015.

[86] Charles S SHERRINGTON. On the proprio-ceptive system, especially in its reflex aspect. *Brain*, 29(4):467–482, 1907.

[87] Uwe Proske and Simon C Gandevia. The proprioceptive senses: their roles in signaling body shape, body position and movement, and muscle force. *Physiological reviews*, 92(4):1651–1697, 2012.

[88] Shao Huang, Weiqiang Wang, Shengfeng He, and Rynson WH Lau. Egocentric hand detection via dynamic region growing. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):10, 2017.

[89] Minjie Cai, Kris Kitani, and Yoichi Sato. Understanding hand-object manipulation by modeling the contextual relationship between actions, grasp types and object attributes. *arXiv preprint arXiv:1807.08254*, 2018.

[90] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050, 2018.

[91] John Weissmann and Ralf Salomon. Gesture recognition for virtual reality applications using data gloves and neural networks. In *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 3, pages 2043–2046. IEEE, 1999.

[92] Deyou Xu. A neural network approach for hand gesture recognition in virtual reality driving training system of spg. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 519–522. IEEE, 2006.

[93] Shahzad Malik, Chris McDonald, and Gerhard Roth. Hand tracking for interactive pattern-based augmented reality. In *Proceedings of the 1st International Symposium on Mixed and Augmented Reality*, page 117. IEEE Computer Society, 2002.

[94] Stefan Reifinger, Frank Wallhoff, Markus Ablassmeier, Tony Poitschke, and Gerhard Rigoll. Static and dynamic hand-gesture recognition for augmented reality applications. In *International Conference on Human-Computer Interaction*, pages 728–737. Springer, 2007.

[95] T. Hollerer and T. Lee. Handy ar: Markerless inspection of augmented reality objects using fingertip tracking. In *2007 11th IEEE International Symposium on Wearable Computers(ISWC)*, volume 00, pages 1–8, 10 2007.

[96] V. I. Pavlovic, T. S. Huang, and R. Sharma. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 19:677–695, 07 1997.

[97] Siddharth S Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, 2015.

[98] Claudio Castellini, Tatiana Tommasi, Nicoletta Noceti, Francesca Odone, and Barbara Caputo. Using object affordances to improve object recognition. *IEEE Transactions on Autonomous Mental Development*, 3(3):207–215, 2011.

[99] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016.

[100] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[101] N. Patricia and B. Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1442–1449, 2014.

[102] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

[103] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015.

[104] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[105] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[106] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[107] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*, pages 3281–3288. IEEE, 2011.

[108] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012.

[109] Xiaofeng Ren and Chunhui Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3137–3144. IEEE, 2010.

[110] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1949–1957, 2015.

[111] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015.

[112] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007.

[113] Matt Huenerfauth and Hernisa Kacorri. Release of experimental stimuli and questions for evaluating facial expressions in animations of american sign language. In *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel, The 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland*, 2014.

[114] German H Flores and Roberto Manduchi. Weallwalk: An annotated dataset of inertial sensor time series from blind walkers. *ACM Transactions on Accessible Computing (TACCESS)*, 11(1):4, 2018.

[115] Hernisa Kacorri, Sergio Mascetti, Andrea Gerino, Dragan Ahmetovic, Hironobu Takagi, and Chieko Asakawa. Supporting orientation of people with visual impairment: Analysis of large scale usage data. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 151–159. ACM, 2016.

[116] Roberta L Klatzky, Brian McCloskey, Sally Doherty, James Pellegrino, and Terence Smith. Knowledge about hand shaping and knowledge about objects. *Journal of motor Behavior*, 19(2):187–213, 1987.

[117] René Gilster, Constanze Hesse, and Heiner Deubel. Contact points during multidigit grasping of geometric objects. *Experimental brain research*, 217(1):137–151, 2012.

[118] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, volume 3, 2016.

[119] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[120] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

[121] Kristen Shinohara and Josh Tenenberg. A blind person's interactions with technology. *Communications of the ACM*, 52(8):58–66, 2009.

[122] Samuel White, Hanjie Ji, and Jeffrey P Bigham. Easysnap: real-time audio feedback for blind photography. In *Adjunct proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 409–410. ACM, 2010.

[123] Marynel Vázquez and Aaron Steinfeld. An assisted photography framework to help visually impaired users properly aim a camera. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(5):25, 2014.

[124] PhotoVoice. Sensory photography, 2019.

[125] Partho Bhowmick. Blind with camera school of photography, 2019.

[126] Dustin Adams, Tory Gallagher, Alexander Ambard, and Sri Kurniawan. Interviewing blind photographers: design insights for a smartphone application. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, page 54. ACM, 2013.

[127] Cynthia L Bennett, Martez E Mott, Edward Cutrell, Meredith Ringel Morris, et al. 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*, page 76. ACM, 2018.

[128] Touch sight: Camera for the blind., 2008.

[129] Dustin Adams and Sri Kurniawan. A blind-friendly photography application for smartphones. *SIGACCESS Access. Comput.*, (108):12–15, January 2014.

[130] Andrea Gerino, Lorenzo Picinali, Cristian Bernareggi, Nicolò Alabastro, and Sergio Mascetti. Towards large scale evaluation of novel sonification techniques for non visual shape exploration. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 13–21. ACM, 2015.

[131] Sergio Mascetti, Lorenzo Picinali, Andrea Gerino, Dragan Ahmetovic, and Cristian Bernareggi. Sonification of guidance data during road crossing for people with visual impairments or blindness. *International Journal of Human-Computer Studies*, 85:16–26, 2016.

[132] Larry D Rosen, Kelly Whaling, L Mark Carrier, Nancy A Cheever, and J Rokkum. The media and technology usage and attitudes scale: An empirical investigation. *Computers in Human Behavior*, 29(6):2501–2511, 2013.

[133] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. Exploring machine teaching for object recognition with the crowd. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages LBW0279:1–LBW0279:6, New York, NY, USA, 2019. ACM.

[134] Thomas J Palmeri and Isabel Gauthier. Visual object understanding. *Nature Reviews Neuroscience*, 5(4):291, 2004.

[135] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y Zhao. Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 225–236. ACM, 2016.

[136] Hernisa Kacorri, Sergio Mascetti, Andrea Gerino, Dragan Ahmetovic, Valeria Alampi, Hironobu Takagi, and Chieko Asakawa. Insights on assistive orientation and mobility of people with visual impairment based on large-scale longitudinal data. *ACM Trans. Access. Comput.*, 11(1):5:1–5:28, March 2018.

[137] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[138] Kyungjun Lee, Abhinav Shrivastava, and Hernisa Kacorri. Hand-priming in object localization for assistive egocentric vision. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, page to appear. IEEE, 2020.

[139] Kyungjun Lee, Abhinav Shrivastava, and Hernisa Kacorri. Leveraging hand-object interactions in assistive egocentric vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[140] Abhinav Shrivastava and Abhinav Gupta. Contextual priming and feedback for faster r-cnn. In *European Conference on Computer Vision*, pages 330–348. Springer, 2016.

[141] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[142] Javier Ribera, David Guera, Yuhao Chen, and Edward J Delp. Locating objects without bounding boxes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6489, 2019.

[143] Zheming Zuo, Longzhi Yang, Yonghong Peng, Fei Chao, and Yanpeng Qu. Gaze-informed egocentric action recognition for memory aid systems. *IEEE Access*, 6:12894–12904, 2018.

[144] Endel Tulving and Daniel L Schacter. Priming and human memory systems. *Science*, 247(4940):301–306, 1990.

[145] Gagan S Wig, Scott T Grafton, Kathryn E Demos, and William M Kelley. Reductions in neural activity underlie behavioral components of repetition priming. *Nature neuroscience*, 8(9):1228, 2005.

[146] Yingfang Meng, Xiaohong Ye, and Brian D Gonsalves. Neural processing of recollection, familiarity and priming at encoding: Evidence from a forced-choice recognition paradigm. *Brain research*, 1585:72–82, 2014.

[147] tephen E Palmer. The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3:519–526, 1975.

[148] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007.

[149] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *2009 IEEE Conference on computer vision and Pattern Recognition*, pages 1271–1278. IEEE, 2009.

[150] Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, page 273, USA, 2003. IEEE Computer Society.

[151] Kevin P Murphy, Antonio Torralba, and William T Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *Advances in neural information processing systems*, pages 1499–1506, 2004.

[152] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[153] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 702–709, June 2012.

[154] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[155] Yukun Zhu, Raquel Urtasun, Ruslan Salakhutdinov, and Sanja Fidler. segdeepm: Exploiting segmentation and context in deep neural networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4703–4711, 2015.

[156] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2359–2367, 2017.

[157] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[158] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13991–14002, 2019.

[159] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7980–7989, 2019.

[160] Dragan Ahmetovic, Daisuke Sato, Uran Oh, Tatsuya Ishihara, Kris Kitani, and Chieko Asakawa. Recog: Supporting blind people in recognizing personal objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[161] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[162] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[163] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[164] Rich Caruana. Multitask learning: A knowledge-based source of inductive bias. In *ICML*, 1993.

[165] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[166] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[167] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA, 2019. Association for Computing Machinery.

[168] Hernisa Kacorri, Utkarsh Dwivedi, Sravya Amancherla, Mayanka K. Jha, and Riya Chanduka. Incluset: A data surfacing repository for accessibility datasets. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, page 26–28, 2020.

[169] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353, June 2012.

[170] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401 – 411, 2017.

[171] Kyungjun Lee, Daisuke Sato, Saki Asakawa, Hernisa Kacorri, and Chieko Asakawa. Pedestrian detection with wearable cameras for the blind: A two-way perspective. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[172] Nelson Daniel Troncoso Aldas, Sooyeon Lee, Chonghan Lee, Mary Beth Rosson, John M Carroll, and Vijaykrishnan Narayanan. Aiguide: An augmented reality hand guidance application for people with visual impairments. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–13, 2020.

[173] Lookout by Google. Use lookout to explore your surroundings, 2021.

[174] Digit-Eyes. Identify and organize your world, 2021.

[175] Xiaoyang Zhang, Harshit Vadodaria, Na Li, Kyoung-Don Kang, and Yao Liu. A smartphone thermal temperature analysis for virtual and augmented reality. *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 301–306, 2020.

[176] Andrea McIntosh, Safwat Hassan, and Abram Hindle. What can android mobile app developers do about the energy consumption of machine learning? *Empirical Software Engineering*, 24(2):562–601, 2019.

[177] Jian Xu, Syed Masum Billah, Roy Shilkrot, and Aruna Balasubramanian. Darkreader: Bridging the gap between perception and reality of power consumption in smartphones for blind users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 96–104, New York, NY, USA, 2019. Association for Computing Machinery.

[178] Gang Luo and Eli Peli. Use of an augmented-vision device for visual search by patients with tunnel vision. *Investigative ophthalmology & visual science*, 47(9):4152–4159, 2006.

[179] Enrico Tanuwidjaja, Derek Huynh, Kirsten Koa, Calvin Nguyen, Churen Shao, Patrick Torbett, Colleen Emmenegger, and Nadir Weibel. Chroma: a wearable augmented-reality solution for color blindness. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pages 799–810, 2014.

[180] Alex D Hwang and Eli Peli. An augmented-reality edge enhancement application for google glass. *Optometry and vision science: official publication of the American Academy of Optometry*, 91(8):1021, 2014.

[181] Yuhang Zhao, Sarit Szpiro, and Shiri Azenkot. Foresee: A customizable head-mounted vision enhancement system for people with low vision. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 239–249, 2015.

[182] Martin Grayson, Anja Thieme, Rita Marques, Daniela Massiceti, Ed Cutrell, and Cecily Morrison. A dynamic ai system for extending the capabilities of blind people. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2020.

[183] S. Seneviratne, Y. Hu, T. Nguyen, G. Lan, S. Khalifa, K. Thilakarathna, M. Hassan, and A. Seneviratne. A survey of wearable devices and challenges. *IEEE Communications Surveys Tutorials*, 19(4):2573–2620, 2017.

[184] NuEyes. Creating inclusive experiences through technology, 2021.

[185] Karst M.P. Hoogsteen, Sjoukje A. Osinga, Bea L.P.A. Steenbekkers, and Sarit F.A. Szpiro. Functionality versus inconspicuousness: Attitudes of people with low vision towards ost smart glasses. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA, 2020. Association for Computing Machinery.

[186] Vuzix Blade. Augment reality (ar) smart glasses for the consumer, 2020.

[187] Guimei Cao, Xuemei Xie, Wenzhe Yang, Quan Liao, Guangming Shi, and Jinjian Wu. Feature-fused ssd: Fast detection for small objects. In *Ninth International Conference on Graphic and Image Processing (ICGIP 2017)*, volume 10615, pages 381–388. SPIE, 2018.

[188] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[189] Jonggi Hong, Jaina Gandhi, Ernest Essuah Mensah, Farnaz Zeraati, Ebrima Jarjue, Kyungjun Lee, and Hernisa Kacorri. Blind users accessing their training images in teachable object recognizers. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '22, New York, NY, USA, 2022. Association for Computing Machinery.

[190] World Health Organization. Who director-general's opening remakrs at the media briefing on covid-19, 2020.

[191] Kyungjun Lee, Daisuke Sato, Saki Asakawa, Chieko Asakawa, and Hernisa Kacorri. Accessing passersby proxemic signals through a head-worn camera: opportunities and limitations for the blind. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–15, 2021.

[192] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[193] Ross Mead, Amin Atrash, and Maja J. Mataric. Recognition of spatial dynamics for predicting social interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, HRI '11, page 201–202, New York, NY, USA, 2011. Association for Computing Machinery.

[194] Leslie A Hayduk and Steven Mainprize. Personal space of the blind. *Social Psychology Quarterly*, pages 216–223, 1980.

[195] Michele A. Williams, Caroline Galbraith, Shaun K. Kane, and Amy Hurst. "just let the cane hit it": How the blind and sighted see navigation differently. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '14, page 217–224, New York, NY, USA, 2014. Association for Computing Machinery.

[196] Stacy M. Branham, Ali Abdolrahmani, William Easley, Morgan Scheuerman, Erick Ronquillo, and Amy Hurst. "is someone there? do they have a gun": How visual information about others can improve personal safety management for blind individuals. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '17, page 260–269, New York, NY, USA, 2017. Association for Computing Machinery.

[197] World Health Organization. Advice for public, 2020.

[198] World Health Organization. Social distancing, quarantine, and isolation: Keep your distance to slow the spread, 2020.

[199] Lighthouse Guild. How people who are blind or have low vision can safely practice social distancing during covid-19, 2020.

[200] American Foundation for the Blind. Flatten inaccessibility study: Executive summary, 2020.

[201] Martina Feierabend, Hans-Otto Karnath, and Jörg Lewald. Auditory space perception in the blind: Horizontal sound localization in acoustically simple and complex situations. *Perception*, 48(11):1039–1057, 2019.

[202] Jillian Rickly, Nigel Halpern, Marcus Hansen, Scott McCabe, and John Fellenor. Covid-19: The effects of isolation and social distancing on people with vision impairment. 2020.

[203] Subeida Ahmed, Harshadha Balasubramanian, Simone Stumpf, Cecily Morrison, Abigail Sellen, and Martin Grayson. Investigating the intelligibility of a computer vision system for blind users. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 419–429, New York, NY, USA, 2020. Association for Computing Machinery.

[204] Martin Grayson, Anja Thieme, Rita Marques, Daniela Massiceti, Ed Cutrell, and Cecily Morrison. A dynamic ai system for extending the capabilities of blind people. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–4, New York, NY, USA, 2020. Association for Computing Machinery.

[205] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. "i am uncomfortable sharing what i can't see": Privacy concerns of the visually impaired with camera based assistive applications. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.

[206] A. Stangl, K. Shiroma, B. Xie, K. R. Fleischmann, and D. Gurari. Visual content considered private by people who are blind. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020.

[207] Bing Li, J Pablo Munoz, Xuejian Rong, Jizhong Xiao, Yingli Tian, and Aries Arditi. Isana: wearable context-aware indoor assistive navigation with obstacle avoidance for the blind. In *European Conference on Computer Vision*, pages 448–462. Springer, 2016.

[208] Annika Muehlbradt, Varsha Koushik, and Shaun K Kane. Goby: A wearable swimming aid for blind athletes. In *Proceedings of the 19th international acm sigaccess conference on computers and accessibility*, pages 377–378, 2017.

[209] Nicoletta Noceti, Luca Giuliani, Joan Sosa-Garciá, Luca Brayda, Andrea Trucco, and Francesca Odone. Designing audio-visual tools to support multisensory disabilities. In *Multimodal Behavior Analysis in the Wild*, pages 79–102. Elsevier, 2019.

[210] Lauren Thevin, Carine Briant, and Anke M Brock. X-road: virtual reality glasses for orientation and mobility training of people with visual impairments. *ACM Transactions on Accessible Computing (TACCESS)*, 13(2):1–47, 2020.

[211] Hong Liu, Jun Wang, Xiangdong Wang, and Yueliang Qian. isee: obstacle detection and feedback system for the blind. In *Adjunct Proceedings of the 2015 ACM International*

Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers, pages 197–200, 2015.

[212] Mohamed L. Mekhalfi, Farid Melgani, Abdallah Zeggada, Francesco G.B. [De Natale], Mohammed A.-M. Salem, and Alaa Khamis. Recovering the sight to blind people in indoor environments with smart technologies. *Expert Systems with Applications*, 46:129 – 138, 2016.

[213] R. Agarwal, N. Ladha, M. Agarwal, K. K. Majee, A. Das, S. Kumar, S. K. Rai, A. K. Singh, S. Nayak, S. Dey, R. Dey, and H. N. Saha. Low cost ultrasonic smart glasses for blind. In *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 210–213, 2017.

[214] Troy McDaniel, Sreekar Krishna, Vineeth Balasubramanian, Dirk Colbry, and Sethuraman Panchanathan. Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind. In *2008 IEEE International Workshop on Haptic Audio visual Environments and Games*, pages 13–18. IEEE, 2008.

[215] Seita Kayukawa, Keita Higuchi, João Guerreiro, Shigeo Morishima, Yoichi Sato, Kris Kitani, and Chieko Asakawa. Bbeep: A sonic collision avoidance system for blind travellers and nearby pedestrians. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[216] Sreekar Krishna, Greg Little, John Black, and Sethuraman Panchanathan. A wearable face recognition system for individuals with visual impairments. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, pages 106–113. ACM, 2005.

[217] Lakshmi Gade, Sreekar Krishna, and Sethuraman Panchanathan. Person localization using a wearable camera towards enhancing social interactions for individuals with visual impairment. In *Proceedings of the 1st ACM SIGMM international workshop on Media studies and implementations that help improving access to disabled users*, pages 53–62, 2009.

[218] Sethuraman Panchanathan, Shayok Chakraborty, and Troy McDaniel. Social interaction assistant: a person-centered approach to enrich social interactions for individuals with visual impairments. *IEEE Journal of Selected Topics in Signal Processing*, 10(5):942–951, 2016.

[219] Cecily Morrison, Edward Cutrell, Martin Grayson, Anja Thieme, Alex Taylor, Geert Roumen, Camilla Longden, Sebastian Tschiatschek, Rita Faia Marques, and Abigail Sellen. Social sensemaking with ai: Designing an open-ended ai experience with a blind child. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[220] Brigitte RoÈder, Wolfgang Teder-SaÈlejaÈrvi, Anette Sterr, Frank RoÈsler, Steven A Hillyard, and Helen J Neville. Improved auditory spatial tuning in blind humans. *Nature*, 400(6740):162–166, 1999.

[221] Albert Postma, Sander Zuidhoek, Matthijs L Noordzij, and Astrid ML Kappers. Differences between early-blind, late-blind, and blindfolded-sighted people in haptic spatial-configuration learning and resulting memory traces. *Perception*, 36(8):1253–1265, 2007.

[222] Amandine Afonso, Alan Blum, Christian Jacquemin, Michel Denis, and Brian FG Katz. A study of spatial cognition in an immersive virtual audio environment: Comparing blind and blindfolded individuals. Georgia Institute of Technology, 2005.

[223] Carla Tinti, Mauro Adenzato, Marco Tamietto, and Cesare Cornoldi. Visual experience is not necessary for efficient survey spatial cognition: evidence from blindness. *Quarterly journal of experimental psychology*, 59(7):1306–1328, 2006.

[224] The Pallets Projects. Flask: web development, one drop at a time, 2020.

[225] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[226] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.

[227] Nicolas Gourier, Daniela Hall, and James L Crowley. Estimating face orientation from robust detection of salient facial features. In *ICPR International Workshop on Visual Observation of Deictic Gestures*. Citeseer, 2004.

[228] Rupal Patel and Deb Roy. Teachable interfaces for individuals with dysarthric speech and severe physical disabilities. In *Proceedings of the AAAI Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 40–47, 1998.

[229] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[230] Heng Yu, Eshed Ohn-Bar, Donghyun Yoo, and Kris M Kitani. Smartpartnet: Part-informed person detection for body-worn smartphones. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1103–1112. IEEE, 2018.

[231] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 792–800, 2018.

[232] Wallace T Cleaves and Russell W Royal. Spatial memory for configurations by congenitally blind, late blind, and sighted adults. *Journal of Visual Impairment & Blindness*, 1979.

[233] Yuhang Zhao, Elizabeth Kupferstein, Brenda Veronica Castro, Steven Feiner, and Shiri Azenkot. Designing ar visualizations to facilitate stair navigation for people with low vision. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pages 387–402, 2019.

[234] Seita Kayukawa, Tatsuya Ishihara, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. Guiding blind pedestrians in public spaces by understanding walking behavior of nearby pedestrians. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(3), September 2020.

[235] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009.

[236] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[237] Shiri Azenkot, Richard E Ladner, and Jacob O Wobbrock. Smartphone haptic feedback for nonvisual wayfinding. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 281–282, 2011.

[238] Jaime Sánchez and Matías Espinoza. Audio haptic videogaming for navigation skills in learners who are blind. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 227–228, 2011.

[239] Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Leah Findlater, and Jon Froehlich. Soundwatch: Exploring smartwatch-based deep learning approaches to support sound awareness for deaf and hard of hearing users. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA, 2020. Association for Computing Machinery.

[240] Yan Yan, Elisa Ricci, Gaowen Liu, and Nicu Sebe. Egocentric daily activity recognition via multitask clustering. *IEEE Transactions on Image Processing*, 24(10):2984–2995, 2015.

[241] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1069–1078, 2021.

[242] Aira. Aira ends support for horizon, 2020.

[243] Tousif Ahmed, Patrick Shaffer, Kay Connelly, David Crandall, and Apu Kapadia. Addressing physical safety, security, and privacy for people with visual impairments. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 341–354, 2016.

[244] Zekeriya Erkin, Martin Franz, Jorge Guajardo, Stefan Katzenbeisser, Inald Lagendijk, and Tomas Toft. Privacy-preserving face recognition. In *International symposium on privacy enhancing technologies symposium*, pages 235–253. Springer, 2009.

[245] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, pages 620–636, 2018.

[246] Yuhang Zhao, Sarit Szpiro, Lei Shi, and Shiri Azenkot. Designing and evaluating a customizable head-mounted vision enhancement system for people with low vision. *ACM Transactions on Accessible Computing (TACCESS)*, 12(4):1–46, 2019.

[247] Steven Goodman, Dhruv Jain, Jon Froehlich, Brock Craft, and Leah Findlater. Social tensions with head-mounted displays for accessibility. 2019.

[248] Leah Findlater, Steven Goodman, Yuhang Zhao, Shiri Azenkot, and Margot Hanley. Fairness issues in ai systems that augment sensory abilities. *arXiv preprint arXiv:1908.07333*, 2019.

[249] Stacy M Branham, Ali Abdolrahmani, William Easley, Morgan Scheuerman, Erick Ronquillo, and Amy Hurst. Is someone there? do they have a gun: How visual information about others can improve personal safety management for blind individuals. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 260–269. ACM, 2017.

[250] Mohammad Kianpisheh, Franklin Mingzhe Li, and Khai N. Truong. Face recognition assistant for people with visual impairments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3), September 2019.

[251] Mat D Duerden and Peter A Witt. The impact of direct and indirect experiences on the development of environmental knowledge, attitudes, and behavior. *Journal of environmental psychology*, 30(4):379–392, 2010.

[252] Julie Rico and Stephen Brewster. Usable gestures for mobile interfaces: evaluating social acceptability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 887–896. ACM, 2010.

[253] David Ahlström, Khalad Hasan, and Pourang Irani. Are you comfortable doing that?: acceptance studies of around-device gestures in and for public settings. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, pages 193–202. ACM, 2014.

[254] Chuang-Wen You, Ya-Fang Lin, Elle Luo, Hung-Yeh Lin, and Hsin-Liu Cindy Kao. Understanding social perceptions towards interacting with on-skin interfaces in public. In *Proceedings of the 23rd International Symposium on Wearable Computers*, pages 244–253. ACM, 2019.

[255] Gerard McAtamney and Caroline Parker. An examination of the effects of a wearable display on informal face-to-face communication. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 45–54. ACM, 2006.

[256] Katrin Wolf, Albrecht Schmidt, Agon Bexheti, and Marc Langheinrich. Lifelogging: You're wearing a camera? *IEEE Pervasive Computing*, 13(3):8–12, 2014.

[257] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In

Proceedings of the 32nd annual ACM conference on Human factors in computing systems, pages 2377–2386. ACM, 2014.

[258] Qianli Xu, Michal Mukawa, Liyuan Li, Joo Hwee Lim, Cheston Tan, Shue Ching Chia, Tian Gan, and Bappaditya Mandal. Exploring users' attitudes towards social interaction assistance on google glass. In *Proceedings of the 6th Augmented Human international conference*, pages 9–12. ACM, 2015.

[259] Valentin Schwind, Jens Reinhardt, Rufat Rzayev, Niels Henze, and Katrin Wolf. On the need for standardized methods to study the social acceptability of emerging technologies. In *CHI'18 Workshop on (Un) Acceptable!?!-Re-thinking the Social Acceptability of Emerging Technologies*, volume 6, 2018.

[260] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, pages 419–429. International World Wide Web Conferences Steering Committee, 2015.

[261] Norene Kelly. The wear scale: Development of a measure of the social acceptability of a wearable device. 2016.

[262] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[263] Utkarsh Dwivedi, Jaina Gandhi, Raj Parikh, Merijke Coenraad, Elizabeth Bonsignore, and Hernisa Kacorri. Exploring machine teaching with children. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–11. IEEE, 2021.

[264] Dhruv Jain, Khoa Huynh Anh Nguyen, Steven M. Goodman, Rachel Grossman-Kahn, Hung Ngo, Aditya Kusupati, Ruofei Du, Alex Olwal, Leah Findlater, and Jon E. Froehlich. Protosound: A personalized and scalable sound recognition system for deaf and hard-of-hearing users. In *CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022.