Revisiting Blind Photography in the Context of Teachable Object Recognizers

Preprint · August 2019

DOI: 10.1145/3308561.3353799

5 authors, including:

Kyungjun Lee
University of Maryland, College Park
8 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Jonggi Hong
University of Maryland, College Park
7 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Hernisa Kacorri
University of Maryland, College Park
39 PUBLICATIONS **210** CITATIONS
SEE PROFILE

Revisiting Blind Photography in the Context of Teachable Object Recognizers

Kyungjun Lee, Jonggi Hong, Simone Pimento, Ebrima Jarjue, Hernisa Kacorri
University of Maryland College Park, USA
kjlee@cs.umd.edu, {jhong12, spimento, ebjarjue, hernisa}@umd.edu

ABSTRACT

For people with visual impairments, photography is essential in identifying objects through remote sighted help and image recognition apps. This is especially the case for teachable object recognizers, where recognition models are trained on user's photos. Here, we propose real-time feedback for communicating the location of an object of interest in the camera frame. Our audio-haptic feedback is powered by a deep learning model that estimates the object center location based on its proximity to the user's hand. To evaluate our approach, we conducted a user study in the lab, where participants with visual impairments ($N = 9$) used our feedback to train and test their object recognizer in vanilla and cluttered environments. We found that very few photos did not include the object (2% in the vanilla and 8% in the cluttered) and the recognition performance was promising even for participants with no prior camera experience. Participants tended to trust the feedback even though they know it can be wrong. Our cluster analysis indicates that better feedback is associated with photos that include the entire object. Our results provide insights into factors that can degrade feedback and recognition performance in teachable interfaces.

Author Keywords

visual impairments; sonification; hand; object recognition

CCS Concepts

•**Human-centered computing** → **Accessibility technologies**; •**Computing methodologies** → **Computer vision**;

INTRODUCTION

Object recognition is one of the daily challenges that people with visual impairments face [17]. 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 [32]. For this reason, beyond adhesive Braille labels [49] or other ad hoc organizing systems [30], people have been quick to adopt camera-based technologies that either use remote sighted help (*e.g.*, [14, 10, 7]) or leverage pre-trained image recognition models (*e.g.*, [56, 25, 53]). Proper aiming of the camera is critical for both cases as it can impact the effectiveness of these solutions.

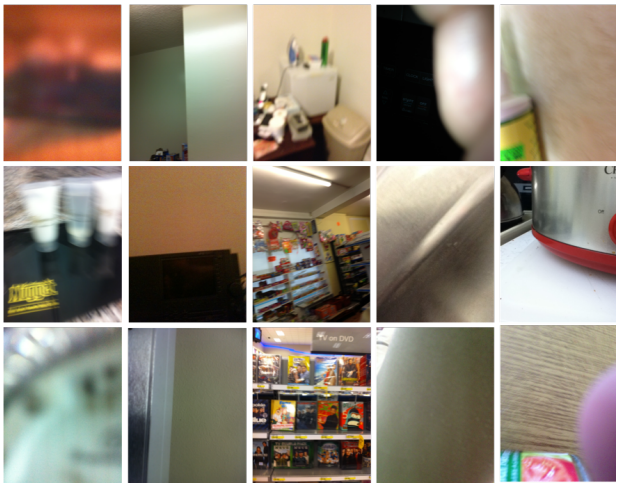


Figure 1: Examples of challenging photos taken by people with visual impairments sent to Vizwiz crowdworkers. About 28% of all photos are categorized as unanswerable [26].

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 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 crowdworkers try to provide feedback or guidance for better camera aiming [8]. 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 [61], given that 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 [32, 50, 5]; 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 [15, 59, 31, 54, 55, 8], it has not been evaluated for object recognition, especially when training machine learning algorithms.

ASSETS'19, October 28–30, 2019, Pittsburgh, PA, USA

ISBN 978-1-4503-6676-2/19/10…$15.00

DOI: https://doi.org/10.1145/3308561.3353799

In this paper, 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 [35] 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 paper'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.

RELATED WORK

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 photography 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 and Nonvisual Feedback

Blind photography is not a new concept. Research in this area includes efforts around the world on teaching photography to people with visual impairments [43, 13]; understanding their photography needs [2, 11, 60]; building accessible photography applications [1, 27, 3, 4]; providing remote sighted help [14, 53, 10, 12]; and, more closely related to this work, obtaining better-quality photos [15, 59, 31, 54, 55, 8, 61, 39].

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 1 presents representative examples from 2010 to 2019 focusing mainly on photos for identifying objects [15, 59, 31, 61] while broadening it to other work helping with the quality of photos for reading barcode or text [39], identifying faces [59, 8, 39, 60], and capturing scenes [54, 55]. As shown in Table 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 [55, 39, 61]. Instead, photos deemed as high quality are extracted automatically from video streams.

Table 1: Comparison of several prior applications addressing camera manipulation for people with visual impairments.

	[15]	[59, 31]	[54, 55]	[8]	[61]	[39]	[60]
Focus							
Object	•	•			•		
Barcode/Text						•	
Face		•		•		•	•
Any			•				
Feedback							
Sound	•	•	•			•	
Verbal		•	•	•			•
Haptic		•		•			
None			•		•	•	
Information							
Face Count							•
Proximity	•					•	
Directions		•	•	•			
Source							
Human	•						
Machine	•	•	•	•	•	•	•

When guidance is provided, it often includes verbal instructions such as left, up, down, right [15, 59, 55, 8] or the number of faces in the camera frame [60] though sonification and haptic feedback have been also explored. Specifically, Bigham *et al.* [15] 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 [39] uses a beeping sound to indicate a good framing. This was contradicted in Vázquez *et al.* [54], 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.* [31] 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.* [24] used a sinusoidal-wave sound and differentiated its frequency to convey spatial information. We consider 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 [38].

Figure 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.

Teachable Object Recognizers

Given the need for fine-grained labels, object recognizers typically rely on the presence of barcodes [29, 40, 39, 6], readable text [44, 39, 6], remote sighted help [14, 10, 12, 7], or large training datasets [53, 56, 19, 25]. 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 [20]. A workaround is to significantly constrain the recognition task to a specific object category such as money readers [45, 6] or a specific user such as teachable object recognizers [51, 32, 50, 5].

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 [32]. However, these approaches are susceptible to camera manipulation challenges faced by people with visual impairments [32, 35]. 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.* [32] as well as Sosa-García and Odone [50] focused on empowering people with visual impairments in training their own object recognizers without sighted help, a focus that this paper 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 [5] and faces of familiar people [5, 6, 39]. 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 [21, 58, 39] or faces in scene description [5, 6, 39]. 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.

HAND-GUIDED OBJECT LOCALIZATION FOR FEEDBACK

Informed by prior work [35], 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 3(b), 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

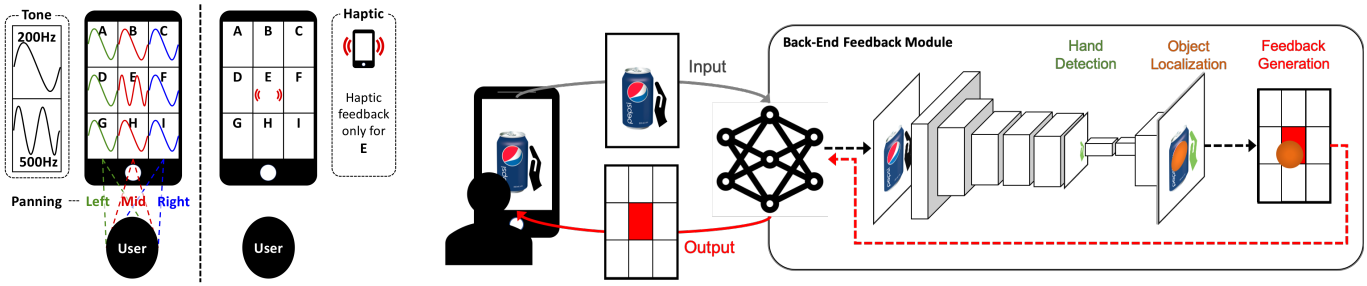
We first train a hand segmentation model that identifies image pixels that correspond to the user's hands in the camera frame. We use a FCN-8s neural network architecture [36] and train it with the publicly available egocentric datasets (a total of $9,241$ training examples): EgoHands [9], GTEA [23], GTEA Gaze+ [22], Intel Egocentric [46], and TEgO [35]. All datasets focus on human interactions from first-person point-of-view: their examples in Figure 2 and a detailed comparison in [35]. 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 [37]. We re-train the remaining layers with Gaussian heatmap blob annotations for object centers ($6,239$ training examples), shown in Figure 2. These annotations were shown to be more robust than pinpointing location coordinates [42, 37]. The hyperparameters for fine-tuning were the same as those noted in the hand segmentation model above.

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.



(a) Our nonvisual feedback mechanism.

(b) The architecture of our real-time feedback approach for better camera framing.

Figure 3: Our testbed sends an image to the server, receives an estimated location, and triggers the corresponding feedback.

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 conveying limited spatial information as highlighted in Vázquez *et al.* [54], 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 3(a). We piloted the stereophonic sound with a beep tone and a sinusoidal wave sound widely used in nonvisual shapes exploration for blind people [24]. 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 [59, 8], 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 3(a), 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.

STUDY AND DATA COLLECTION

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.* [32], 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.

Table 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

Participants

We recruited nine participants with visual impairments from our local community (IRB #1255427-2) each compensated at 55–80 (μ: 63.3, σ: 7.1). As shown in Table 2, all participants identified as legally blind; five of them reported having some light perception. They were all female age 29–69 (μ: 56, σ: 15). 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 shows participants' use and attitudes toward technology potentially affecting our task, based on the Rosen *et al.* [47] 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 other five participants were: KNFB reader [44], BeMyEyes [10], Aira [7], and SeeingAI [6].

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 3(b). The estimated location is based on the image pixels classified with the highest probability as the center of

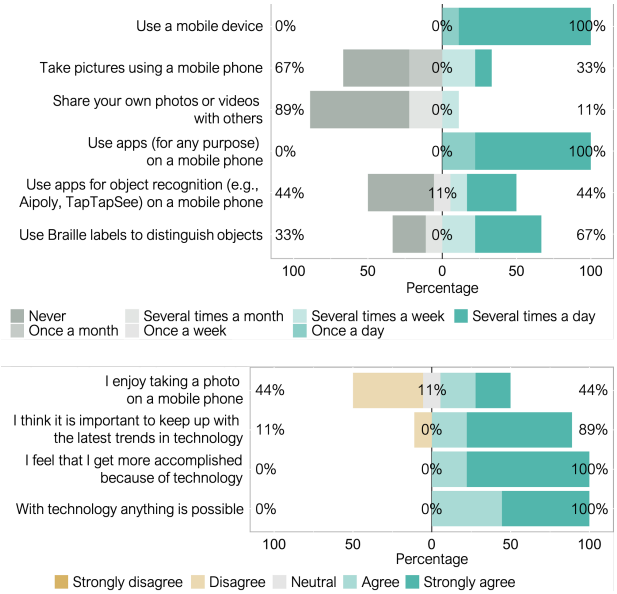


Figure 4: Technology experience and attitude responses. All participants have smartphones; more than half are using apps for object recognition; and all are positive about technology.

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.

Object Stimuli

As shown in Figure 5(a), we use three objects for practice. The grill salt and mountain dew can are selected from the TEgO dataset [35] 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 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.* keychains or artisanal products that may or may not have readable texts or recognizable labels), researchers [32, 50] 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.* [32] (Figure 5(b)). Due to product changes, labels for some of the stimuli such as *k-cups*,

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

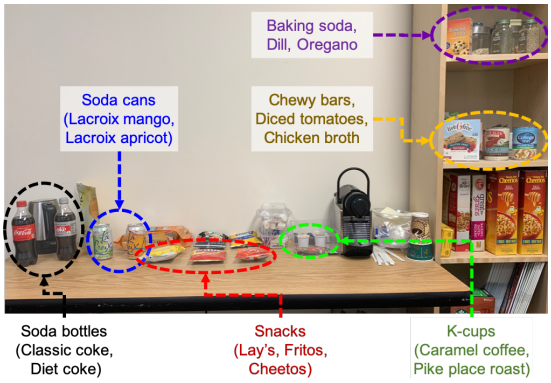


Figure 6: The wild test setting in the cluttered environment.

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.

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 7) as in Kacorri *et al.* [32]. 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 6). We control for lighting conditions across participants by having both environments indoors without natural light.

Study Procedure

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 technology. 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.

Figure 7: Vanilla environment and the camera manipulation across participants in train mode.

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 [32]:

- *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.

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.* [32], we used the same recognition model based on Google's Inception V3 [52] pre-trained on ImageNet [48] 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.* [32].

Similar to Kacorri *et al.* [32], 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.

Observations and Findings

Promising performance for those with no camera experience. Figure 8 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.* [32], 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. 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

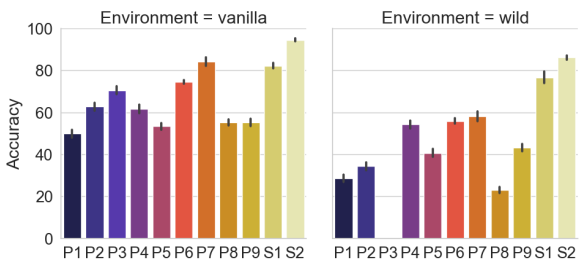


Figure 8: Average model accuracy per participant.

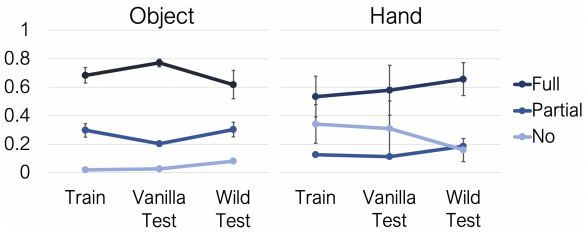


Figure 9: Proportion of photos with fully, partially, or not included objects and hands. Errors bars show variance among participants, who were able to fully capture the object in 60–80% of their photos and include their hand in more than 50%.

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 9). Last, it could be due to changes in photo variations; variations in vanilla photos seem to be different from those in the wild (Table 3).

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 9), we adopted the codes in Hong *et al.* [28] based on the four dimensions that humans generalize across for visual recognition [41]: size, location, viewpoint, and illumination. Specifically, as shown in Table 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).

Table 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

Observations and Findings

Limited number of photos without the object of interest. Figure 9 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 (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 9 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 10); all participants agreed that this is feasible, and most of them 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."*

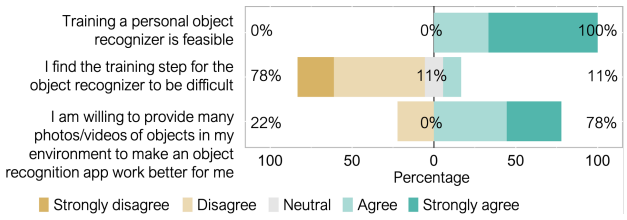


Figure 10: Post-study feedback. Most participants thought training is not difficult and they were willing to provide photos to train and improve the performance of an object recognizer.

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 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.

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.* [34] 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 [57, 34, 33]. 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 [16]. As shown in Figure 11, 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 [33].

Observations and Findings

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 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 13 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

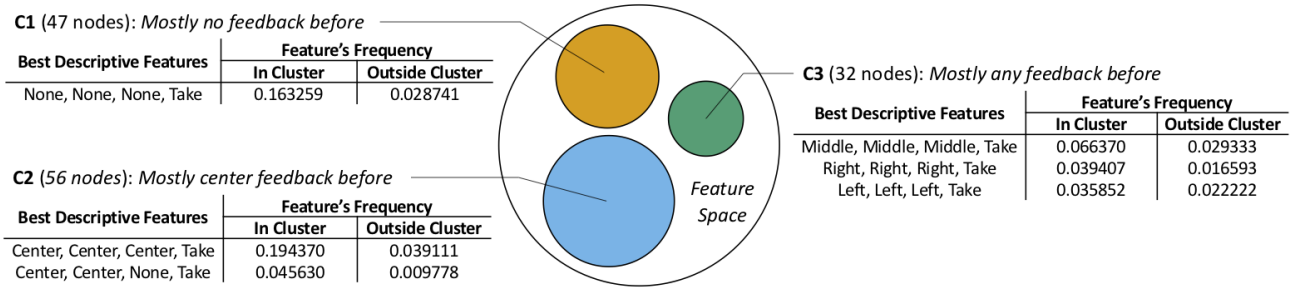
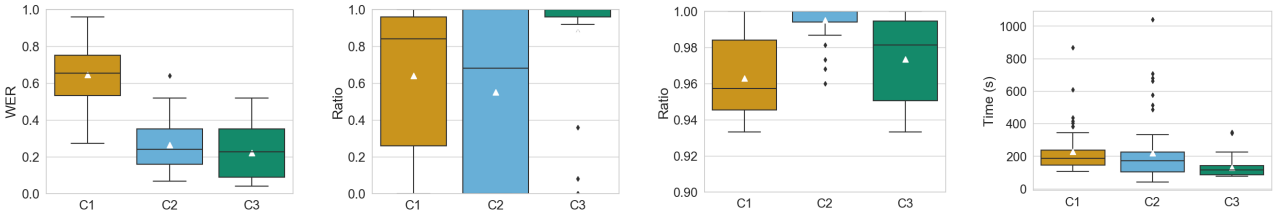


Figure 11: By representing the stream of interactions in a (participant, object) pair as a unique node, we find three distinct clusters and their descriptive features that are most responsible for this cluster formation.



(a) Feedback WER across pairs. (b) % of photos including hands. (c) % of photos with full object. (d) Time spent in training.

Figure 12: Comparing clusters across different interaction characteristics during training.

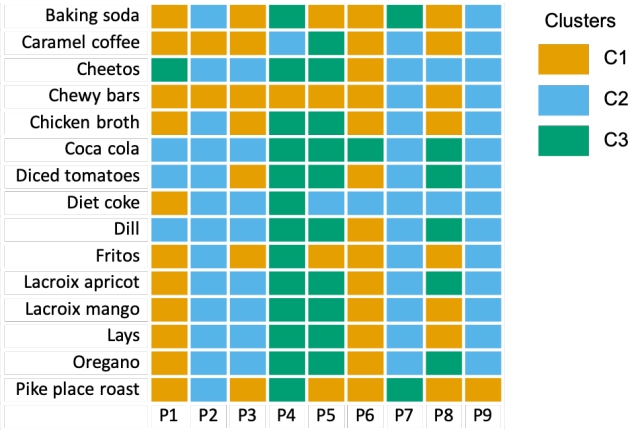


Figure 13: Cluster distribution across participant-object pairs.

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 [1], σ: 0.24). As shown in Figure 12(a), we also provide pair-level WERs across the clusters.

[1]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.

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 12). 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.

DISCUSSION

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

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.* [32] 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.

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 [50, 32], 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 [28]; 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.

CONCLUSIONS

This work 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. We believe that this work has been instrumental for better understanding some of the challenges that users with visual impairments may face when interacting with such teachable interfaces, and hope that it contributes to future work in this direction.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for insightful comments on earlier drafts of this paper. This work is supported by NIDILRR (Award: #90REGE0008). Jonggi Hong and Hernisa Kacorri are supported in part by NSF (Award: #1816380). Ebrima Jarjue is supported in part by CRA-W in partnership with AccessComputing (DREU).

REFERENCES

[1] 2008. Touch Sight: Camera for the blind. (2008). http://www.yankodesign.com/2008/08/13/thiscamera-is-outta-sight/

[2] Dustin Adams, Tory Gallagher, Alexander Ambard, and Sri Kurniawan. 2013. Interviewing blind photographers: design insights for a smartphone application. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 54.

[3] Dustin Adams and Sri Kurniawan. 2014. A blind-friendly photography application for smartphones. *ACM SIGACCESS Accessibility and Computing* 108 (2014), 12–15.

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

[5] Envision AI. 2018. Enabling vision for the blind. (2018). https://www.letsenvision.com

[6] Seeing AI. 2017. A free app that narrates the world around you. (2017). https://www.microsoft.com/en-us/seeing-ai

[7] Aira. 2017. Your Life, Your Schedule, Right Now. (2017). https://aira.io

[8] Jan Balata, Zdenek Mikovec, and Lukas Neoproud. 2015. BlindCamera: Central and Golden-ratio Composition for Blind Photographers. In *Proceedings of the Mulitimedia, Interaction, Design and Innnovation*. ACM, 8.

[9] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. 2015. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision*. 1949–1957.

[10] BeMyEyes. 2015. Bringing sight to blind and low-vision people. (2015). http://www.bemyeyes.org

[11] Cynthia L Bennett, Martez E Mott, Edward Cutrell, Meredith Ringel Morris, and others. 2018. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 76.

[12] BeSpecular. 2016. Let blind people see through your eyes. (2016). https://www.bespecular.com

[13] Partho Bhowmick. 2019. Blind with Camera School of Photography. (2019). http://www.blindwithcameraschool.org/

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

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

[16] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

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

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

[19] CamFind. 2013. Search the physical world. (2013). http://camfindapp.com

[20] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie. 2016. 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*. 1153–1162.

[21] Digit-Eyes. 2010. Identify and organize your world. (2010). http://www.digit-eyes.com

[22] Alireza Fathi, Yin Li, and James M Rehg. 2012. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*. Springer, 314–327.

[23] Alireza Fathi, Xiaofeng Ren, and James M Rehg. 2011. Learning to recognize objects in egocentric activities. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*. IEEE, 3281–3288.

[24] Andrea Gerino, Lorenzo Picinali, Cristian Bernareggi, Nicolò Alabastro, and Sergio Mascetti. 2015. 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*. ACM, 13–21.

[25] Talking Goggles. 2013. A camera with speech. (2013). http://www.sparklingapps.com/goggles

[26] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. *arXiv preprint arXiv:1802.08218* (2018).

[27] Susumu Harada, Daisuke Sato, Dustin W Adams, Sri Kurniawan, Hironobu Takagi, and Chieko Asakawa. 2013. 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*. ACM, 2127–2136.

[28] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2019. 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)*. ACM, New York, NY, USA, Article LBW0279, 6 pages. DOI: http://dx.doi.org/10.1145/3290607.3312873

[29] i.d. mate. 2018. Talking bar code scanners. (2018). http://www.envisionamerica.com/store

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

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

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

[33] Hernisa Kacorri, Sergio Mascetti, Andrea Gerino, Dragan Ahmetovic, Valeria Alampi, Hironobu Takagi, and Chieko Asakawa. 2018. Insights on Assistive Orientation and Mobility of People with Visual Impairment Based on Large-Scale Longitudinal Data. *ACM Trans. Access. Comput.* 11, 1, Article 5 (March 2018), 28 pages. DOI: http://dx.doi.org/10.1145/3178853

[34] Hernisa Kacorri, Sergio Mascetti, Andrea Gerino, Dragan Ahmetovic, Hironobu Takagi, and Chieko Asakawa. 2016. 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*. ACM, 151–159.

[35] Kyungjun Lee and Hernisa Kacorri. 2019. Hands Holding Clues for Object Recognition in Teachable Machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

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

[38] Sergio Mascetti, Lorenzo Picinali, Andrea Gerino, Dragan Ahmetovic, and Cristian Bernareggi. 2016. Sonification of guidance data during road crossing for people with visual impairments or blindness. *International Journal of Human-Computer Studies* 85 (2016), 16–26.

[39] OrCam MyEye. 2019. User Guide. (2019). https://www.orcam.com/wp-content/uploads/2018/11/orcam-myeye2-8-2-user-guide_en_141118.pdf

[40] Opticon. 2018. Handheld Scanner. (2018). http://www.opticonusa.com/products/handheld-solutions

[41] Thomas J Palmeri and Isabel Gauthier. 2004. Visual object understanding. *Nature Reviews Neuroscience* 5, 4 (2004), 291.

[42] Tomas Pfister, James Charles, and Andrew Zisserman. 2015. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 1913–1921.

[43] PhotoVoice. 2019. Sensory Photography. (2019). https://photovoice.org/methodologyseries/method_04/index.htm

[44] KNFB Reader. 2018. Access to print materials. (2018). http://www.knfbreader.com

[45] NantMobile Money Reader. 2017. Instantly recognizes currency and speaks the denomination. (2017). https://nantmobile.com

[46] Xiaofeng Ren and Chunhui Gu. 2010. Figure-ground segmentation improves handled object recognition in egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 3137–3144.

[47] Larry D Rosen, Kelly Whaling, L Mark Carrier, Nancy A Cheever, and J Rokkum. 2013. The media and technology usage and attitudes scale: An empirical investigation. *Computers in Human Behavior* 29, 6 (2013), 2501–2511.

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[49] Kristen Shinohara and Josh Tenenberg. 2009. A blind person's interactions with technology. *Commun. ACM* 52, 8 (2009), 58–66.

[50] Joan Sosa-García and Francesca Odone. 2017. "Hands On" Visual Recognition for Visually Impaired Users. *ACM Transactions on Accessible Computing (TACCESS)* 10, 3 (2017), 8.

[51] Jeremi Sudol. 2013. *LookTel—Computer Vision Applications for the Visually Impaired*. Ph.D. Dissertation. UCLA.

[52] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[53] TapTapSee. 2012. Mobile camera application designed specifically for the blind and visually impaired iOS users. (2012). `http://www.taptapseeapp.com`

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

[55] Marynel Vázquez and Aaron Steinfeld. 2014. An assisted photography framework to help visually impaired users properly aim a camera. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 5 (2014), 25.

[56] Aipoly Vision. 2016. Sight for Blind & Visually Impaired. (2016). `http://aipoly.com`

[57] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y Zhao. 2016. Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 225–236.

[58] WayAround. 2018. The smart assistant for people who are blind. (2018). `https://www.wayaround.com`

[59] Samuel White, Hanjie Ji, and Jeffrey P Bigham. 2010. EasySnap: real-time audio feedback for blind photography. In *Adjunct proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 409–410.

[60] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2018. A Face Recognition Application for People with Visual Impairments: Understanding Use Beyond the Lab. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 215.

[61] Yu Zhong, Pierre J Garrigues, and Jeffrey P Bigham. 2013. 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*. ACM, 20.