People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges

Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, Chieko Asakawa
Carnegie Mellon University
Pittsburgh, USA
hkacorri@cmu.edu, kkitani@cs.cmu.edu, jbigham@cs.cmu.edu, chiekoa@cs.cmu.edu

ABSTRACT
Blind people often need to identify objects around them, from packages of food to items of clothing. Automatic object recognition continues to provide limited assistance in such tasks because models tend to be trained on images taken by sighted people with different background clutter, scale, viewpoints, occlusion, and image quality than in photos taken by blind users. We explore personal object recognizers, where visually impaired people train a mobile application with a few snapshots of objects of interest and provide custom labels. We adopt transfer learning with a deep learning system for user-defined multi-label *k*-instance classification. Experiments with blind participants demonstrate the feasibility of our approach, which reaches accuracies over 90% for some participants. We analyze user data and feedback to explore effects of sample size, photo-quality variance, and object shape; and contrast models trained on photos by blind participants to those by sighted participants and generic recognizers.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (*e.g.*, HCI)

Author Keywords
blind; accessibility; photographs; object recognition; computer vision

INTRODUCTION
Blind people have been quick to adopt assistive technologies for identifying objects such as text readers, barcode readers, color readers, and crowd-powered object recognition applications. However, there are still many limitations that hold back their use for more independent living. For example, a barcode reader, *e.g.*, i.d. mate [9], may cost up to a few thousand dollars, and requires a readable barcode and an updated product database. Text readers, such as KNFBReader [15], work best on flat surfaces and printed materials and are thus not applicable for many object recognition tasks. Color readers such as Color Teller [22] typically used to help with outfit matching, can be very sensitive to lighting conditions, forcing the user to memorize color mappings to distinguish between clothes under different illumination. Applications that use the crowd to identify an object, such as BeMyEyes [3], BeSpecular [4] and TapTapSee [21], can obtain high recognition rates but often come with a per demand cost, require an Internet connection, raise privacy concerns [1], and assume crowd availability.

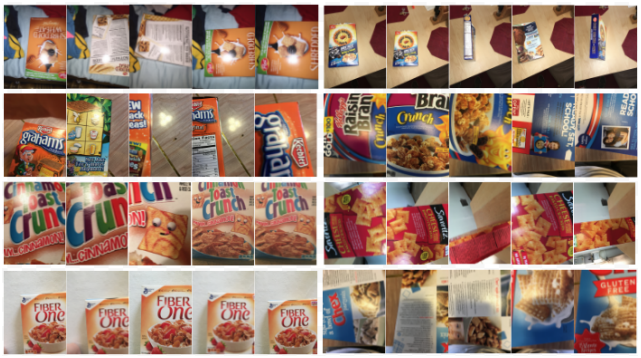

Figure 1. Object instances that participants in our study chose to train their personal object recognizers on. Can you tell which two objects were trained by the same participant? (4,2) (2,2) .rewsnA

On the other hand, the development of assistive technologies that make use of a commodity smartphone's camera and on-board processing overcome many of the challenges listed above. Smartphones have been adopted by many people with visual impairments. Cameras can be used to capture both flat and 3D objects. By doing all processing on the device, the technology is not dependent on an Internet connection and naturally ensures privacy. Based on these observations, we are interested in examining the feasibility (and challenges) of a self-contained image-based object recognition algorithm trained specifically for and by people with visual impairments.

Benefits of Personalization. One of the key conditions for developing a robust real world assistive computer vision system, is to significantly constrain the imaging conditions of the recognition task. One such successful example for people with visual impairment are money readers, which work for a small number of object classes, *e.g.*, bills of different denominations. Conversely, building a super object classifier which can recognize all the possible object instances of interest for all visually impaired people, is not possible with current object recognition technology.

CHI 2017, May 06 - 11, 2017, Denver, CO, USA.

ACM 978-1-4503-4655-9/17/05..$15.00.
DOI: http://dx.doi.org/10.1145/3025453.3025899

Assuming that we cannot train on all the possible object instances of interest to all visually impaired people, we suggest constraining the recognition task through *personalization*. We adopt the concept of a personal object recognizer, where people with visual impairment collect (sequentially) a small sample of photos from their object of interest and provide custom labels to train their personalized application. By limiting the recognition task to a single user, we not only constrain the number of object classes but also reduce the variability between training and testing images. This is natural since such an application would be trained and tested by the same person under similar conditions and subjected similarly to any idiosyncratic characteristics, as shown in Fig. 1.

Feasibility Study. While many blind people use Braille labels to annotate their objects, or some *ad hoc* organizing system, they might not be familiar with the concept of training a personal recognizer though photo taking. As Phase 1 of our study, we provid participants with a description of how our technology works along with some basic photo taking instructions. We solicit suggestions from users on how such technology might be used and ask them to photograph a few snapshots of objects in their homes over a one-week period.

Data collected in Phase 1 are analyzed to estimate the instance distribution for objects of interest, and confirm our intuition that object classes in our problem best fit within the scenario of finer grained sets of labels. We use observations from results in Phase 1 to compile a set of training instructions and invite the participants to our lab for a round of simulated training and testing under controlled conditions.

Experiments from Phase 2 demonstrate the potential of our approach, which reaches accuracies over 90% for some participants. Our error analysis, combined with observations from images and video recordings from Phase 2, uncovers user factors that can degrade the performance of a personal object recognizer and should be prioritized for robustness of a real world application. We also contrast our personal object recognizer with different approaches using recognizers trained by sighted people and a generic object recognizer to confirm the feasibility of our approach.

Contributions. We develop a framework for allowing people with visual impairments to train an image-based personalized object recognition algorithm and show that one can achieve high accuracy by constraining the task through personalization. Our feasibility study shows that personalization through fine-grained object recognition is in fact, a necessity for many people with visual impairments. Furthermore, our experiments with blind participants show that personalized object recognition algorithms have superior performance over generic state-of-the-art object recognition models.

RELATED WORK

Prior work in object recognition for people with visual impairment mainly spans two areas, crowd-powered and automatic object recognition. There are few cases in each area where image classifiers are trained on images captured by blind users. TapTapSee [21], a crowd-powered application, tries to automate the objection recognition process by training models on previously seen images with matching crowdsourced labels, while preserving a human-backed image recognition approach. Aipoly [2], a generic image recognition application, obtains "image–label" pairs from blind users by asking them to teach the application with image and description pairs. In both cases, training is neither immediate nor personalized. Images provided by one user may adversely affect the performance of the model for any other user.

The closest application to our work is LookTel Recognizer [16]. This application allows users to recognize objects given a library of training images. It requires "high-quality, well-framed images with ample lighting" captured with the assistance of sighted people. To our knowledge, our work is the first study of the feasibility of blind people training their own object recognizers. To understand differences due to photos taken by sighted people, we also compare models trained by blind versus sighted participants in our study.

Another line of research involves obtaining higher quality photos from blind users. Crowd-powered applications directly solicits camera positioning guidance from a crowd-worker until a good quality image is obtained. Researchers, *e.g.*, [25], have also investigated automated approaches for extracting good quality information-rich frames from continuous camera video streams. In both cases, a good quality photo is implicitly defined as one that is either interpreted properly by a sighted person or as one that maximizes the recognition from a model trained on images taken by sighted people.

We suspect that a high quality personal object recognizer can work with a looser notion of quality. For our approach, both training and testing images are provided by the same person, under similar conditions and for a bounded number of object classes. As a consequence, the image may not contain the entire object. Instead, an image is considered of good quality if there is enough consistency across training and testing images for any given object class, and there are sufficient discriminative characteristics to limit confusion between classes.

PERSONAL OBJECT RECOGNIZER

State-of-the-art image recognition trains machine learning models that achieve performance comparable to, or even exceeding, human recognition, with performance reported against ImageNet [19], an established benchmark for object classification with millions of images. Typically, thousands of examples are needed to train an image classifier. However, we can use transfer learning by leveraging fully trained recognition models, and adapting them to new image classification tasks using only a few samples. Transfer learning [13] applies knowledge from one model to solve a different but related problem. Here, the labels differ while the marginal distributions of data are related [14].

Baseline. As our baseline object recognizer, we use Inception-v3 network [20] (hereafter, ''Inception"), a state-of-the-art model from Google which achieves a 3.46% top-5 error rate[1] on ImageNet when classifying images into 1000 classes.

[1]Top-5 error rate indicates how often the model fails to give the correct answer in the top 5 guesses.

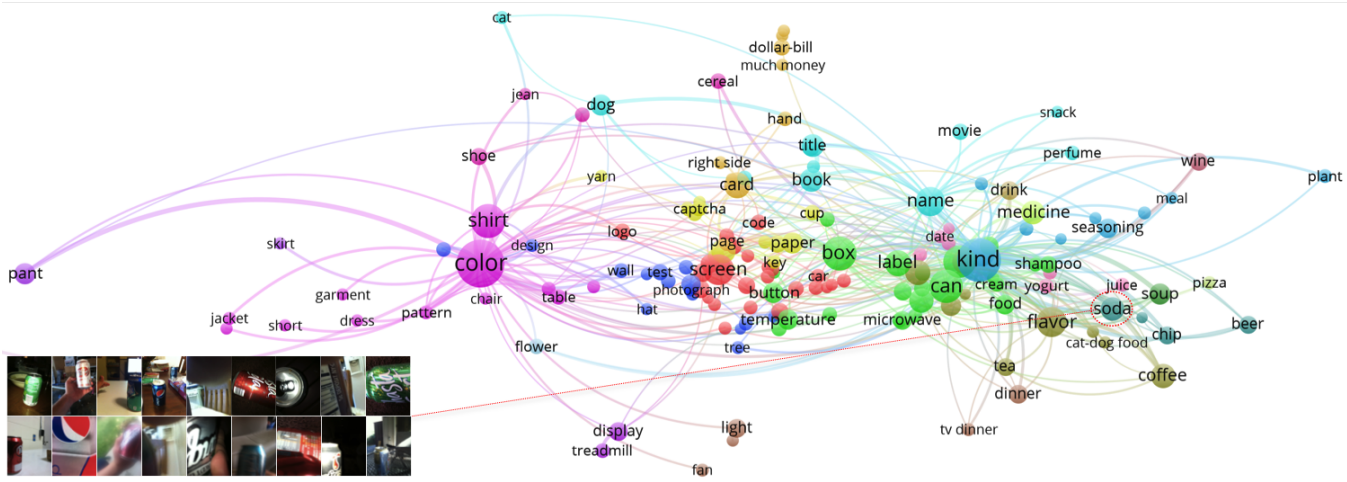

Figure 2. Term co-occurrence network for user questions in VizWiz dataset with sample images for the co-occurring terms 'soda' and 'kind'.

Our classifier. Google released a pre-trained Inception allowing researchers to build higher level machine learning layers. With transfer learning, an Inception model trained on ImageNet can be retrained or 'fine-tuned' for new image classes. The intuition is that lower layers in Inception's pre-trained model have learned generic low-level representations for distinguishing objects. Thus, these generic features can be re-purposed for other recognition tasks. To build a personal object recognizer, we load the pre-trained Inception, replace the top layer with a new (softmax) layer trained on the new user-defined classes. The new layer yields probabilities for each class. We take the class with the highest probability to be the network's classification for a given image. While our experiments process examples in batch mode, in a realistic scenario, photos for a new class would be processed one class at a time by adding a new node to the final layer.

A key challenge to our problem is that the sample size for training has to be limited to a practical small number. Our problem is an instance of *N-way k-shot learning*, where k is a small number of samples, *e.g.*, 1-20, and N is the number of object classes. Our approach follows recent research using a pre-trained Inception on ImageNet for k-shot learning tasks. For example, Vinyals et al. (2016) [24] showed that a pre-trained Inception on ImageNet is a competitive baseline for N-way k-shot learning, outperforming proposed state-of-the-art methods when the label distribution is fine grained. Similarly, prior work in fine-grained image classification, such as the LifeCLEF challenge [7], also used a trained GoogLeNet CNN (a prior version of Inception) on ImageNet and replaced its top layers as in our approach, achieving high performance. Therefore, transferring knowledge using a pre-trained Inception on ImageNet for N-way k-shot classification of a fine-grained dataset is a competitive approach.

We do not know *a priori* potential biases in object categories of interest for people with visual impairment. However, we hypothesize that object classes in our problem will better fit the scenario of fine grained sets of labels as in the previous works [24, 7]. Our intuition is that people with visual impairment are interested in distinguishing between objects with similar shape and texture, difficult to tell by touch. We examine this further in the following sections.

PRELIMINARY INSIGHTS FROM VIZWIZ DATASET

To gain better insights on our recognition task, we analyzed the publicly available VizWiz dataset[2], which includes questions and images from people with visual impairment accompanied by human-backed image recognition and answers provided by a sighted crowd. While an initial analysis classified these questions into identification, description, and reading categories [5], the focus of our analysis is to identify the types of object categories and characteristics being asked as well as their relationships.

Specifically, we apply VOSViewer text mining [23] to VizWiz users' questions, a total of 33,543 transcribed questions. We extract the most relevant terms and create a term co-occurrence network, as shown in Figure 2. There are a total of 163 terms occurring at least in 20 questions. Each term is visualized as a node and co-occurring terms are linked with an edge. Term occurrence and co-occurrence is denoted by node size and edge thickness. Terms with high relevance are grouped together into clusters, and each cluster may be seen as a topic governing the question.

The co-occurence network in Figure 2 allows us to make the observation that blind users often know the general object category such as shirt, bottle, can, soup, soda, coffee, medication but are interested in *specific* characteristics of those objects such as color, kind, flavor, label, brand, and name. This highlights the need of an instance object recognizer over generic image classification.

Zooming into one of the nodes in Figure 2 and examining the photos taken by blind users, we observe large variability not only in the instances of soda cans but also in the background clutter, scale, viewpoints, occlusion, and image quality. The diversity of the images indicates that it may be quite difficult to learn an image classifier that can be robust to such extreme intra-class variance.

[2]http://vizwiz.org/data/

STUDY AND DATA COLLECTION

The VizWiz dataset is rich with images taken by people with visual impairment but does not contain sequential photos of the same object taken by a given user. Thus, we are unable to use VizWiz to assess the feasibility of our approach. To understand the potential of learning with few examples given by blind people, we designed a two phase study. We recruited 8 blind participants (P1-P8) from the local community who were familiar with smartphones and asked them[3] to take photos of objects in their homes (Phase 1) and in our lab (Phase 2).

Table 1 shows participants' demographic information and period of smartphone usage. Fig. 3 reports their media and technology usage, and attitudes potentially affecting our task, based on the Rosen et al. questionnaire [17]. In open-end questions, P1 and P8 reported on home appliances using Braille labels, and P3-P7 reported on objects that are: easy to confusing, frequently used, big enough to accommodate labels, or in new visiting environments. Participants indicated that making Braille labels is time consuming and requires a lot of discipline. When possible, they use alternative strategies for distinguishing between similar objects based on touch (*e.g.* shape and size), sound (*e.g.*, shaking a jar), smell, weight, order (cans stacked based on flavors), and location (different cabinets). When these strategies failed, users would turn to technology, with the following applications and devices mentioned across 8 participants: BeMyEyes [3], BeSpecular [4], CamFind [6], Color Teller [22], Facetime (to ask family members and friends), i.d. mate Galaxy [9], KNFB Reader [15], Opticon [12], Talking Goggles [8], and TapTapSee [21].

ID	*Gender*	*Age*	*Blind (since)*	*Handedness*	*Smartphone*
P1	F	42	birth	right	2014
P2	M	40	birth (light)	right	2015
P3	F	68	birth	right	2008
P4	M	63	birth	left	2009
P5	F	46	birth	right	2016
P6	M	43	birth	right	2006
P7	F	61	birth (light)	right	2010
P8	F	58	10 months	right	2011

Table 1. Participants' demographics and smartphone use period.

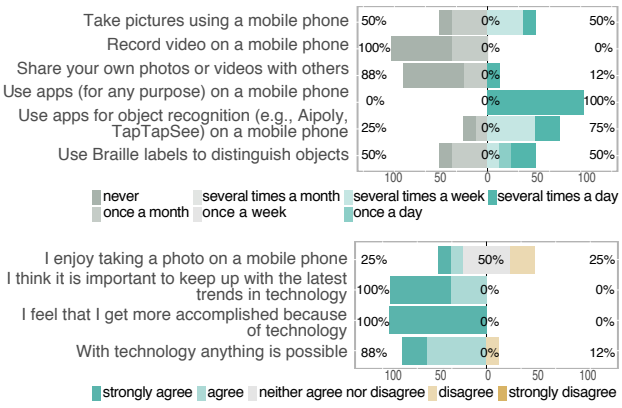

Figure 3. Technology experience and attitude responses.

[3]Under IRBSTUDY2015_00000052

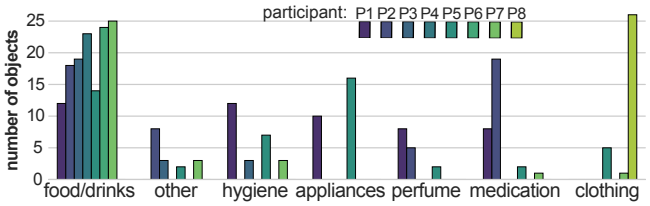

Figure 4. Distribution of object instances chosen by the participants.

Phase 1: In Situ

Over a week period, participants were asked to take photos of up to 50 objects of their choice, spending no more than 2 hours overall. Participants were asked to imagine that they are building a personalized phone application which uses the camera to recognize objects in *their* home. We further explained that the application is personalized in that it will work well only for the small number of objects on which it will be trained. Thus, they would have to choose the top 50 objects they would like their personalized application to recognize. Participants were instructed to label and take at least 5 photos for each of these objects on top of an empty table without surrounding objects. We suggested positioning the phone 8 to 12 inches from the object to ensure that it was within the scope of phone camera. Additional instructions on accessible photo taking were also included in the task description.

The goal of this phase was to allow participants time to process what a personal object recognizer entails before reporting their feedback about this technology, provide an estimate of the target domain with class distribution per classifier, and inform the characteristics and challenges of this task.

Observations and Findings in Phase 1

We received photos of 23 to 50 (average 35.12) distinct objects per participant with a total of 1,543 photos across participants. Figure 4 illustrates the distribution of the object instances that participants chose for training a personal recognizer. We observe that this distribution shares similarities with the term graph extracted from VizWiz users in Figure 2.

Drilling down within object groups, we observe that participants choose object instances typically falling under the same category such as "dove body wash winter" and "summer body wash dove"; and, "seasoned breadcrumbs" and "unseasoned breadcrumbs". Also of interest, we notice that for one of the participants, P8, all objects of interest are t-shirts with different patterns. This confirms our intuition that the classes of objects in our problem best fit within the scenario of finer grained sets of labels.

In our analysis of the *user-defined labels* for the selected objects, we observed:

Preference for personalized metadata. Beyond knowing brand, label, name, color, scent, flavor, as in the VizWiz data in Figure 2, participants would also like to record when and where they obtained the objects, washing instructions, cost, *etc.*. For example, one of the participants labeled some objects as ".... medals: Id race and year", "T-shirts: color, washing, where from and year".

Challenges in assigning labels to objects. Participants had a difficult time assigning labels to objects in their homes. This highlights the limitations of a personal object recognizer, which assumes that during training participants will have knowledge of object labels. For example, participant would label some of their objects as "box of cereal", "another box of cereal", "K cup, unknown flavor", "Pasta sauce, not sure what kind".

At the end of this study phase, we asked the participants to indicate some of their reasoning for choosing their objects. We believe that their responses, shown below, highlight user expectations from this technology:

> *"things I did not wanted to wait for a response", "things that take more time to figure out the other ways, time-based", "things that I thought would be interesting* e.g.*, seasoned breadcrumbs versus non seasoned breadcrumbs", "things that I want to distinguish from others (brand and flavor).* e.g.*, cookies, all of their boxes look similar and I don't want to open before", "things that I lost track of what they are", "cans and bottles that look similar", "things based on pretty much what I was doing those days (spice, cans, food) and a lot of those things are marked in Braille", "things that I would pull out of the draw and ask a family member or use the color identifier but it doesn't work. It was one of the things it would get mixed up easily".*

There are no previous datasets of sequential photos of objects taken by blind people for the purpose of training a classifier. Thus, we examined the *collected images* to identify interesting patterns and user behaviors that could be informative for the task at hand to drive Phase 2 of our study. We observed:

Distinct and consistent training strategies. Participants tend to develop distinct training strategies. Figure 1 illustrates photos of 8 products selected by 7 participants to train their personal object recognizer. Two products, rows 2 and 4 in the second column, are taken by the same participant. While not instructed to do so, participants tried to introduce some variation across the 5 images per object. And their perception of how to produce such variation for training also varied: different viewpoints, distances from the object, rotation, and visible side. However, it is interesting that each participant tended to be consistent given an object shape. If this consistency continues in the testing mode of a personally trained application it could lead to higher recognition rates.

Exaggerated or non-discriminative viewpoints. Some of the training images in the sequence of 5 include exaggerated viewpoints and scale in an attempt to provide a holistic view of the object rather than capture discriminative characteristics, *e.g.*, last photos of cylindrical objects tend to be from the top of the lid with no texture or visual characteristics. This could be an artifact of how training is perceived by the participants. We suspect one of the limitations of this approach to be the fact that images used for training are captured sequentially and in a different mode than testing. Typically, successful machine learning models are trained on images pulled from a similar distribution as in testing. Not only will the training images for a given object have limited background and light variation, since they are taken sequentially, but it could be that these exaggerated training images will not be observed in everyday usage, as noticed in the VizWiz dataset.

Pre-compiled set of instructions. Based on our observations from this phase of the study, we compiled a set of instructions for blind people to train a personal object recognizer. These instructions are used in Phase 2 of the study.

1. *Increase consistency across training and testing.* When taking training photos, imagine how your future self could be holding and taking a photo of that object to identify it.
2. *Limit background clutter.* Lay the object down in a table or any other flat surface. If it is a cylindrical shape it is okay to hold it to stabilize it, as long as your hand does not cover the object.
3. *Ensure that the object is in the camera scope*: feel where the camera is located in the phone and position the camera to the center of the object. Move away from the object upwards while holding the phone parallel to the desk.
4. *Ensure that most of the object is in the camera scope.* Distance the phone from the object relative to the object size (closer for smaller far way for larger).
5. *Obtain more 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.

Phase 2: In Laboratory

We selected a subset of 15 object instances and asked Phase 1 participants to train and test a personal object recognizer in a lab setting. Object instances, shown in Figure 5, fall under the food/drink group which had the highest object occurrence across most participants in Phase 1 of the study. This is shown in Figure 4 and in agreement with the most relevant terms in the VizWiz network in Figure 2. The goal of this study was to collect data that will allow us to assess the feasibility of our approach and explore effects of photo-quality variance, sample size, and object characteristics. Our primary interest was understanding performance variation due to inherent differences in how blind users conceptualize and perform training tasks. We minimized variations in conditions, present in Phase 1, such as background, discriminative characteristics of objects, and lighting. We ensured that participants operated with a common understanding of how such an application would

Figure 5. Phase 2 objects: baking soda, cheetos, chewy bars, chicken broth, coca cola, diced tomatoes, diet coke, dill, fritos, lacroix apricot, lacroix mango, lay's, oregano, pike place roast, pike place roast decaf.

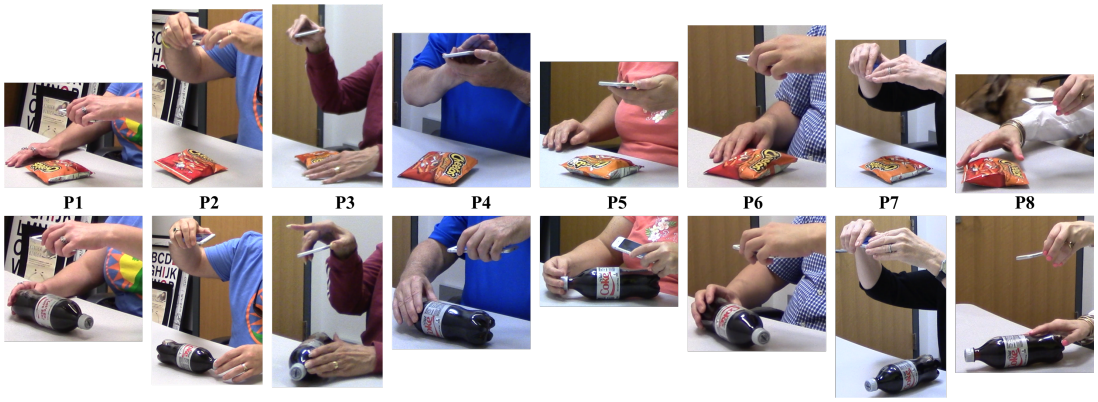

Figure 6. Idiosyncratic object and camera manipulation across participants during training.

Figure 7. Training and testing images from P1 and P5 on "chewy bars" object illustrating distinct training strategies. (Images are uniformly sampled from the available 30 training images.)

work by providing them with the pre-compiled instructions, resulting from Phase 1, and feedback in a practice session with 3 distinct objects from Figure 5. We also controlled for variation in the order and way in which objects are passed to the user, e.g. randomized order and randomized up-side of objects, to simulate real-world scenarios. This allows us to not only compare performance variability across different participants but also to contrast models trained on photos by blind participants to those by sighted participants and generic recognizers.

Phase 2 has two main modes: training and testing. In the simulated training mode, participants were asked to take about 30 sequential photos of each of the 15 objects. In the simulated testing mode, participant were only able to take one photo of each given object to try and identify it. To minimize learning effects, objects were shuffled consistently across participants, in such a way that there were 5 photos for each object and two consecutive photos were never of the same object.

The lab setup was identical in practice, training, and testing modes for all participants. To ensure identical lighting conditions, a room without natural light was chosen. To take the photos, participants used VoiceOver on an iPhone 6 device and were recommended to use the volume buttons as in the practice mode. To draw insightful observations, each session was recorded on video.

In this phase, we also collected data from two sighted people. The first person (S1) is one of the authors of this paper, who is familiar with Inception and the transfer learning approach being adopted, serves merely as an upper baseline to the discriminative power of model for the 15 selected object instances and the background and lighting settings in this study. The strategy here was to collect high quality images, shown in Figure 5, with very limited variability, both during training and testing, while following the pre-compiled instructions as well as visual feedback from the camera. The results from S1 can be used to interpret results from other participants by excluding effects due to the particular task or experiment setting.

The second person (S2) is a sighted female, age 38, who is not familiar with Inception or other image classification tasks and machine learning. S2 serves as a second sighted person for an upper baseline of good quality images and received the same information about this experiment as the participants. Both S1 and S2 followed the practice, training, and testing modes as the blind participants under the same setting.

Observations and Findings in Phase 2

We collected a total of 4,120 photos in training mode and 661 in testing mode from our participants P1-P8. A participant spent on average 65 seconds to take about 30 training photos for an object ($23 - 245$[4], std: 35.2).

When looking at the participants' photos and video recordings of the sessions, we observed:

Presence of user's hand in training images. To take photos, participants used one hand (P1, P3, P5, P6, P8), two hands (P7), or interchanged between two hands (P2, P4) as shown in Figure 6. Excepting P8, all participants tend to include one hand in the training photos to either hold an object, or simply as a reference point to ensure that the object is in the photo. However, for very few participants, this

[4]The peek of 245 seconds was observed for the baking soda since the package was leaking and needed more care when rotating.

behavior was mirrored in the testing photos with P1, P2, and P6 being the most consistent. We expect these differences to be reflected in the performance of the personal object recognizer trained by each participant in our experiments.

Reinforced distinct and consistent training strategies. Our observations in Phase 1 about distinct training strategies were reinforced in Phase 2. With the option of more photos per object, 30 photos compared to 5 in Phase 1, participants introduced user-distinct variations that focused more on the visible face of the object, up side-down rotations, viewpoints, and distances from the object. Figure 7 illustrates the difference between the visible faces variation and distance variation introduced by P1 and P5.

Variation in training unobserved in testing. The exaggerated viewpoints and scale discussed in the observations of Phase 1 seem to hold here too. For example, independently of how the "chewy bars" box was passed to P1 in testing mode, the participant consciously avoided viewpoints from the narrow box opening side, while initially such a viewpoint was included many times towards the end of the training examples. Participants were instructed to increase consistency across training and testing by imagining how their future self would be taking the photos. However, many of them perceived the number of 30 examples as high and often interpreted it as an opportunity for variation and covering of edge cases.

Different training strategies among sighted participants. Interestingly, we also observed a big difference between sighted people S1 and S2. S1, aware that the highest accuracy is achieved with maximal consistency across training and testing images, introduced minimal variance in the training samples and preserved that consistency during testing, *e.g.*, taking photos from the front of the object to cover most of the phone screen. On the other hand, S2 introduced variation in viewpoints, sides, and distance from the object hoping to give the application "a more holistic view" of the object. We suspect that training strategies are affected by the way people perceive machine learning concepts, such as training a recognizer, and can thus make a difference in the classifier performance.

We are interested in gauging willingness of blind users to take the time of training their application and taking photos for this purpose. At the end of the study, we asked the participants to indicate whether training a personal object recognizer is feasible with responses shown in Figure 8. We also solicited comments on possible challenges they faced during simulated training and testing in Phase 2 of the study. We found that participants agreed with the feasibility of training a personal object recognizer from the blind user's perspective. Their main concerns were: knowing whether the photos were good, knowing the area of a package where the label or distinguishing information resides, obtaining feedback from the camera such as lighting conditions and number of photos taken, deciding on the distance between the object and camera lens. One of the participants stated "*the most challenging and most fun is training the person*."

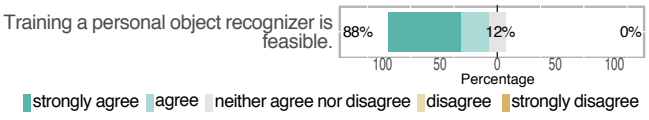

Figure 8. Subjective responses on the feasibility of our approach.

EXPERIMENTS AND RESULTS

To assess the proposed approach, we used data obtained in Phase 2 to build personal object recognizers for each participant (P1-P8 and S1-S2) and report their performance in terms of accurately predicting labels for the corresponding testing images belonging to 15 object classes in Figure 5. The hyperparameters for our models were: 1000 training steps, 0.01 learning rate, 2048-dimensional feature vector, and 299 x 299 input image. Training images were always pulled from the set of photos collected in training mode in Phase 2. Testing images included all images in testing mode, about 75 images per participant. Given that the image data were obtained from a controlled study, to avoid overfitting to characteristics of the data related to task and experiment settings, we did not perform any further fine tuning of the model parameters.

Model Performance

Figure 9 illustrates results from personal object recognizers trained and tested on data from P1-P8 and S1-S2. The number of training images obtained in Phase 2 varied slightly across objects and participants. For more comparable results, we randomly selected 20 images of each object per participant across all experiments. To account for randomness in the selection process and randomness inherent in the training process, we built 10 model attempts per object recognizer and controlled for random seeds for reproducibility throughout all experiments.

For blind participants P1-P8, the personal object recognizers achieved accuracies ranging from 50.7% up to 92% ($\mu = 75.95\%, \sigma = 13.29$). For comparison, a random prediction for a 15-way classification would yield about 7% accuracy, while a model trained on data from S1 a sighted computer scientist with understanding of the underlying algorithms achieved an average accuracy of 99.6%, and a model trained on S2, a sighted person unaware of the underlying methods, achieved an average accuracy of 96.9%.

As expected, S1, taking images with minimal variation across training and setting, achieved almost perfect scores and out-

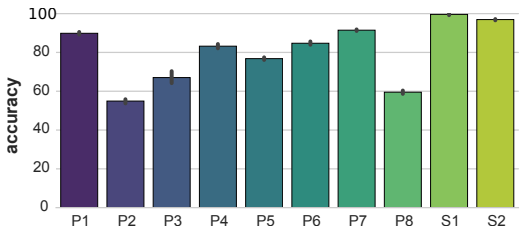

Figure 9. Results of a personal object recognizer trained and tested on images from each participant with 20 samples per object and error bars calculated over 10 random experiment runs.

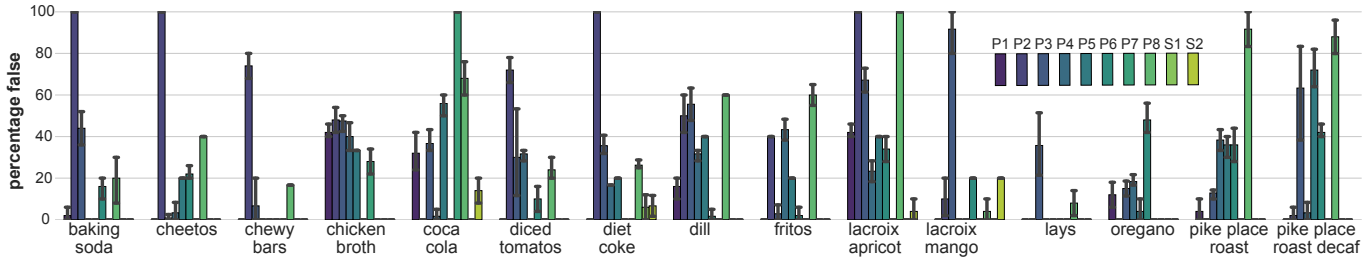

Figure 10. Percentage of misclassified images per object category across all participants indicating that difficult-to-recognize objects are not uniformly distributed across participants. (A 100% indicates that all 5 test images within that category were misclassified and error bars show variation over 10 random experiment runs.)

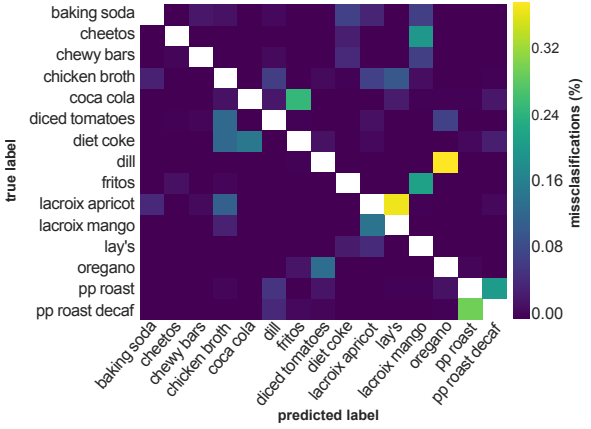

Figure 11. Most mislabeled objects over P1-P8 recognizers share shape and visual similarities. (Percentages are averaged over 10 random experiment runs.)

performed all participants. This serves as an upper baseline for the performance of our approach given the object stimuli, experimental setup of Phase 2, and the hyper parameters of our model. The fact that some of our participants P1 and P8 achieved performances comparable to those of a sighted participant S2 highlights the potentials of our approach.

Error Analysis

While the results on testing accuracy of Figure 9 allowed us to quickly grasp the potential of this technology, we looked deeper for how to improve upon it. In this section, we focus our attention on the errors and attempt to link them with prior observations from the study.

Figure 10 shows the percentage of misclassified images per object category, across all participants, with error bars calculated over the 10 trials. Our first observation here is that objects which are difficult to recognize are not uniformly distributed across participants. For example, focusing on P2 and P8, whose models achieve the lowest accuracies (seen in Fig 9), we observe P2's model consistently misclassifies baking soda, cheetos, diet coke, and lacroix mango, whereas P8's model consistently misclassifies lacroix apricot, pike place roast, and pike place roast decaf.

Examining the training and testing images for P2 and P8 on the highly misclassified objects, we find that their recognizers' performances vary for different reasons. Participant P2's strategy for introducing variation during training tended to be consistent with the variation in testing. The degradation in classifier performance was due to idiosyncratic characteristics of the participant. As shown in Figure 6, P2 tended to hold the camera at a higher distance from the object and with a slight tilt away from the object. As a result, the object is marginally included or excluded in many of the training photos. For example, none of the training data for diet coke included the object in the image frame. Thus, it is critical for a real world application to account and provide feedback for the presence of the object during training, either by detecting the tilt in users phone or by using computer vision techniques.

However, from participant's P8's images among highly misclassified objects, we observed that training images included exaggerated viewpoints, *e.g.*, the last photos of lacroix apricot tend to be from the top of the can and many of the last k-cup photos were taken from the side, while none of the testing photos for those objects where taken from those perspectives. Our models were trained on 20 randomly selected images from the participant's training pool, and we believe that more intelligent sampling techniques for selecting training data can be employed to build more robust models.

Figure 11 illustrates aggregated results on misclassified testing images across participants P1-P8 as a confusion heatmap. We observe that the highest confusions are: lacroix apricot as lacroix mango, dill as oregano, pike place roast decaf as pike place roast, coca cola as diet coke, fritos as lay's, and cheetos as lay's. These are all objects that were intentionally chosen due to shared shape and visual similarities. Interestingly, even for the sighted participant S2, the misclassified pairs were (lacroix apricot, lacroix mango) and (coca cola, diet coke). We suspect that there is room for better performance in our models orthogonal to the quality of photos taken by blind people, *e.g.* additional data, more training steps, and tuning of other parameters.

Effect of Order in Training Examples

Our initial results were obtained from personalized classifiers trained on 20 randomly sampled images per object from the available pool of participant's training images. However, training images for any given object are acquired sequentially from users. We wondered whether training on a batch of photos taken sequentially would affect performance. Further, would it make a difference where this sequential batch is taken from among the entire sequence? This could matter if participants modulate their photo-taking strategies as they manipulate the

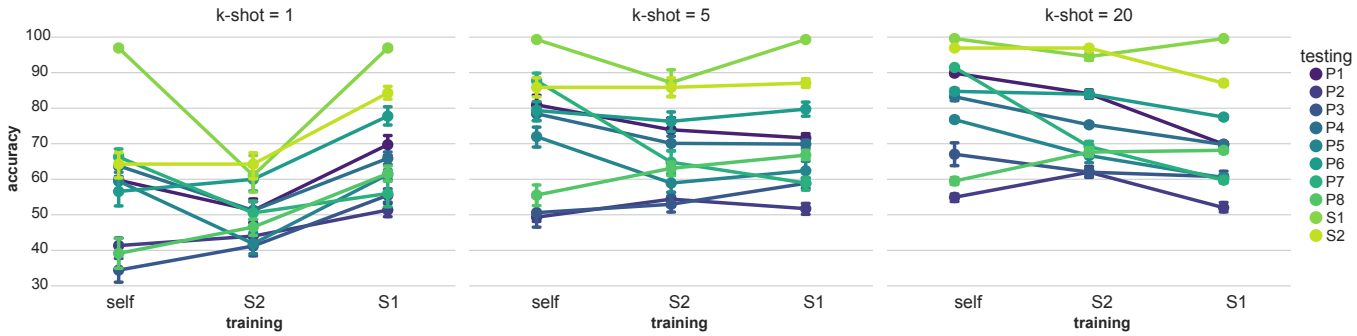

Figure 12. A personal object recognizer trained and tested on images from the same blind participant tends to outperform a recognizer trained on images from a sighted person S1 or S2 when the training size is bigger ($k\_shot = 20$).

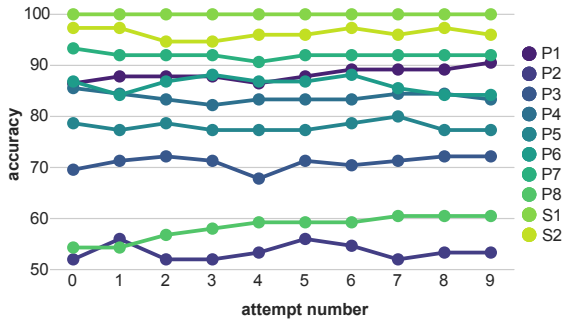

Figure 13. Selecting training samples through a sliding window did not have similar effect on recognizers' accuracy across participants.

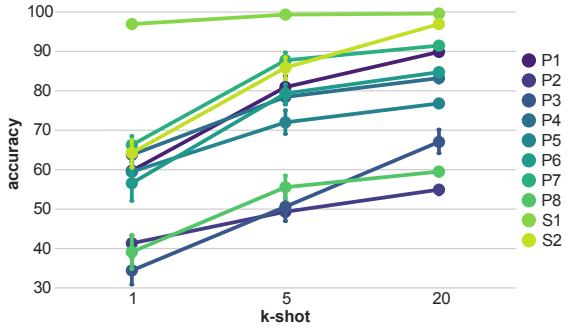

Figure 14. Recognizer's accuracy increases with the sample size, where $k = 1, 5, 20$.

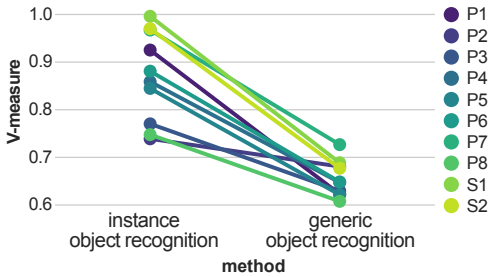

Figure 15. A personal object recognizer achieves higher desirable classification properties than a generic state-of-the-art object recognizer.

objects. To investigate this, we ran a set of experiments where the 20 samples were selected sequentially from the pool of available training images per object. For each participant's recognizer, we ran 10 attempts, where each attempt shifts the selected images by one. For example, in attempt 0, we train on the first 20 images, *i.e.* images $1, \ldots, 20$, in attempt 1, we train on images $2, \ldots, 21$, and so on. On comparing recognizers' performance on random versus sequential training images, we did not observe a consistent effect across all participants P1-P8. Moreover, we did not observe a common pattern when drilling down into achieved performance across participants per attempt in the sequential selection, shown in Figure 13. This may be because many of our stimuli objects were cylindrical or that participants develop different strategies to train their personal object recognizer. As above, we believe that more sophisticated approaches need to be investigated for selecting a good subset of training images.

Effect of Sample Size and Data Augmentation

One issue raised by the participants is how Braille labeling is time consuming and requires a lot of discipline. Even though we observed that participants were able to take, on average, 30 photos of an object over 65 seconds, we explored the potential of the proposed approach for training with highly limited sample sizes of 1 and 5 with k-shot learning [5]. Figure 14 illustrates the model accuracies across participants per value of k. In 1-shot learning, blind participants' recognizers achieved an average accuracy of 52.6% ($23.5\% - 73.3\%, \sigma = 12.7$) with P7's classifiers outperforming those of a sighted person, S2. Across all participants, higher k resulted in better accuracy.

The most minimal improvement was observed for S1, who serves as an upper baseline in these experiments, and for the blind participant P2. For S1, this can be explained by the fact that both training and testing images were high quality and introduced minimal variation by intention. However, for P2 this is as an effect of many training images which exclude the object, as discussed in the section on error analysis. We believe that presence of outlier images (i.e. unrepresentative of the object) in the training data can bias the solution, severely degrading classification performance in k-shot learning.

A common practice to improve the performance of machine learning models is to augment the original training data with

[5]Results in k-shot learning are typically reported for $k = 1, 5, 20$.

new data generated by applying different distortions on the original ones. To demonstrate the potentials of data augmentation for our approach we expanded the effective size of the training data by randomly cropping and scaling up to 20% of the image, and by applying up to 10% brightness to training images of participants P5, P6, and P7. By adding possible variation of the same images we could help the personal object classifier be more robust on distortions that can occur in testing. Initial results show that accuracy might be boosted by up to 10.7% ($\mu = 3.8\%, \sigma = 3.2$).

Contrast to Models Trained on Photos by Sighted People

As discussed in the related work section, LookTel Recognizer, the closest application to our work, requires training on high-quality and well-framed images taken by sighted people in the user's environment. While we cannot directly compare LookTel performance to our approach, we explore the benefits of having a sighted person training a blind user's personal object recognizer. Specifically, we compare the accuracy achieved by models trained and tested on images by the same person to the accuracy of models trained on images by S1 or S2 but tested on images from all all participants.

Results, shown in Figure 12, show that models trained on a single example ($k = 1$) perform best when the training image is taken by a sighted person, with better performance using S1's recognizer. However, for larger training samples, a personal object recognizer trained and tested on images from the same blind user tends to outperform a recognizer trained on images from a sighted person. Specifically, for $k = 20$ we observe that the accuracy of predicted labels drops for 6 out of 8 blind participants, with P2 and P8 as exceptions.

Contrast to Generic Image Recognition

While not a straightforward comparison, we wanted to explore how a personal object recognizer trained to recognize object instances performs relative to a generic image recognition model, the method used by most mobile applications for the blind. Since a generic recognizer is not trained on user-defined labels, an accuracy score cannot be calculated directly. Thus we cannot directly contrast its performance with our approach. However, we can compare 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. For example, assume a blind user provides a photo to AppX, an imaginary mobile application with a fully trained Inception model, and receives the top-1 predicted label. The user is interested in distinguishing between the 'pike place roast' and the 'pike place roast decaf k-cups', shown in Figure 5 and has noticed that AppX recognizes the first as 'Petri dish' and the second as 'bottlecap'. If AppX consistently identifies 'pike place roast' as 'Petri dish' and the decaffeinated as 'bottlecap', then the user could learn a $1-1$ mapping and still benefit from AppX despite the confusion in actual object labels. Though such a solution is not scalable, its utility in limited scenarios led us to contrast it with our approach. For this comparison, we adopt the V-measure [18] metric from cluster evaluation, which is independent of the absolute values of the labels. Thus, it measures the agreement of two independent label assignment strategies on the same data. First, we calculate V-measures for each of the participants' recognizers in Fig 9 using true labels for testing images and the corresponding predicted labels as inputs. We average V-measures across the 10 experiment runs for each participant. To obtain the V-measure from a fully-trained Inception model, not adapted to the participants, we use true labels for testing images per participant against Inception's predicted labels belonging to 1000 possible classes (distinct from user-defined labels).

Figure 15 illustrates the obtained V-measures contrasted per each participant. We observe that for all the participants, our approach achieves higher V-measures, where a 100% accuracy would correspond to a V-measure 1.0.

CONCLUSION AND FUTURE WORK

We have presented a method that allows people with visual impairments to train a personal object recognition algorithm to differentiate between everyday objects specific to the user. By doing so, we have provided a solution for a very practical need for people with visual impairments – instance level object recognition. Additionally, by leveraging the constraints on the image data distribution through personalization, we were able to create image classifiers that can adapt state-of-the-art generic classifiers for user specific tasks using only a small number of examples.

Based on our observations and findings, our future work will explore user interface design and computer vision methods to increase robustness in training. This will involve automating and providing user feedback on some of the pre-complied set of instructions *e.g.* indicating detected background clutter based on image analysis and a tilting phone based on the accelerometer. We found that one of the main reasons for performance degradation is the absence of the object of interest from the training examples due to challenges in photo-taking by blind users. Based on our observation of the users tendency to use their hand as a guiding point of reference for the camera, we will incorporate elements from our prior work on first-person activity recognition [11] to select training examples where the object appears in the vicinity of the user's hand. Moreover, we will investigate intelligent sampling techniques for selecting representative and diverse subsets of training examples from discriminative viewpoints.

While there are many parameters that can still be evaluated (*e.g.*, incremental model learning, extreme illumination changes, video versus images), the more important remaining issue is one of scalability over long periods of time. Although we have shown success for a moderately sized dataset, what happens as the number of objects increase over time to hundreds or thousands of labels? We believe that this work has been instrumental in understanding the use of personalized object classifiers for a short duration of time (*i.e.*, 2 hour time frame) and hope that it serves as an impetus for large-scale long-term evaluation (*e.g.*, based on remote usage data [10]) to validate the longevity and utility of such technologies.

ACKNOWLEDGMENTS

This work has been supported by Shimizu Corporation. Kris Kitani is supported in part by JST CREST.

REFERENCES

1. Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. 2015. Privacy Concerns and Behaviors of People with Visual Impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3523–3532. http://doi.acm.org/10.1145/2702123.2702334
2. Aipoly. 2016. Vision through artificial intelligence. (2016). http://aipoly.com/
3. BeMyEyes. 2016. Lend you eyes to the blind. (2016). http://www.bemyeyes.org/
4. BeSpecular. 2016. Let blind people see through your eyes. (2016). https://www.bespecular.com/
5. Erin L Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P Bigham. 2013. 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*. ACM, 1225–1236.
6. CamFind. 2016. Search the physical world. (2016). http://camfindapp.com/
7. Julien Champ, Titouan Lorieul, Maximilien Servajean, and Alexis Joly. 2015. A comparative study of fine-grained classification methods in the context of the lifeclef plant identification challenge 2015. In *CLEF 2015*, Vol. 1391.
8. Talking Goggles. 2016. A camera with speech. (2016). http://www.sparklingapps.com/goggles/
9. i.d. mate. 2016. Talking bar code scanners. (2016). http://www.envisionamerica.com/store
10. 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 *18th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM.
11. 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*.
12. Opticon. 2016. Handheld Scanner. (2016). http://www.opticonusa.com/products/handheld-solutions/
13. Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2010), 1345–1359.
14. Novi Patricia and Barbara Caputo. 2014. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1442–1449.
15. KNFB Reader. 2016. Access to print materials. (2016). http://www.knfbreader.com/
16. LookTel Recognizer. 2016. Instantly recognize everyday objects. (2016). http://www.looktel.com/recognizer
17. 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.
18. Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*, Vol. 7. 410–420.
19. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.
20. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567* (2015).
21. TapTapSee. 2016. Mobile camera application designed specifically for the blind and visually impaired iOS users. (2016). http://www.taptapseeapp.com/
22. Color Teller. 2016. The Talking Color Identifier. (2016). http://www.brytech.com/colorteller/
23. Nees Jan Van Eck and Ludo Waltman. 2011. Text mining and visualization using VOSviewer. *ISSI Newsletter* 7, 3 (2011), 50–54.
24. Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. In *In Proceedings of the Conference on Neural Information Processing Systems (NIPS'16)*.
25. 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.