HHS Public Access

Author manuscript

Proc SIGCHI Conf Hum Factor Comput Syst. Author manuscript; available in PMC 2020 February 10.

Published in final edited form as:

Proc SIGCHI Conf Hum Factor Comput Syst. 2019 May ; 2019: . doi:10.1145/3290605.3300566.

Hands Holding Clues for Object Recognition in Teachable Machines

Kyungjun Lee,
Department of Computer Science, University of Maryland, College Park, MD, USA

Hernisa Kacorri
College of Information Studies, University of Maryland, College Park, MD, USA

Abstract

Camera manipulation confounds the use of object recognition applications by blind people. This is exacerbated when photos from this population are also used to train models, as with teachable machines, where out-of-frame or partially included objects against cluttered backgrounds degrade performance. Leveraging prior evidence on the ability of blind people to coordinate hand movements using proprioception, we propose a deep learning system that jointly models hand segmentation and object localization for object classification. We investigate the utility of hands as a natural interface for including and indicating the object of interest in the camera frame. We confirm the potential of this approach by analyzing existing datasets from people with visual impairments for object recognition. With a new publicly available egocentric dataset and an extensive error analysis, we provide insights into this approach in the context of teachable recognizers.

Keywords

blind; object recognition; hand; egocentric; k-shot learning

1 INTRODUCTION

Object recognition apps making use of built-in cameras on mobile or wearable devices have gained in popularity among blind users as they help with access to the visual world. Typically, they provide solutions employing sighted help through crowdsourcing (*e.g.*, Aira

kjlee@cs.umd.edu.

CCS CONCEPTS

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

ACM Reference Format:
Kyungjun Lee and Hernisa Kacorri. 2019. Hands Holding Clues for Object Recognition in Teachable Machines. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300566

[7] and BeMyEyes [9]) or machine learning with pre-trained image recognition models (*e.g.*, SeeingAI [5] and Aipoly Vision [67]). They have also employed hybrid solutions, falling back to sighted help when computer vision models fail (*e.g.*, TapTapSee [65]). Camera manipulation, however, remains a challenge for blind users [34, 36, 66, 73]. In all cases, the effectiveness of these solutions is confounded by blurred images, viewpoints with low discriminative characteristics, cluttered backgrounds, low saliency, and more importantly *partially included* or *out-of-frame* objects of interest.

Human-powered applications, especially those employing video services, can provide near-real-time guidance for better camera aiming. However, they often come with other limitations such as the need for crowd availability and Internet connection, per demand cost, and privacy concerns [3, 61, 71]. Such limitations can be resolved with pre-trained computer vision models using the device's onboard processing, though many of the camera manipulation challenges remain difficult to resolve. For example, a computer vision app, which is typically trained on well-framed photos of sighted people, may not be able to recognize the object of interest when given a partially included object against a cluttered background. In the case of object recognizers with teachable machines [35, 36, 62], where machine learning models are trained and personalized by blind users, having objects partially included or even out of the camera frame is a primary factor limiting performance [36]. To overcome these challenges, we are interested in examining hands as a natural interface for including and indicating the object of interest in the camera frame (as shown in Fig. 1).

Why hand-detection in object recognition.

When exploring the feasibility of teachable object recognizers for the blind, Kacorri *et al.,* 2017 [36] observed that users' hands tend to be present in many of the images, either holding an object or serving as a reference point to place the object in the photo. We speculate that they were leveraging proprioception [51, 60], the perception of body and limb position, to coordinate their hand movements: one hand holding and adjusting the camera to the location of the other hand that was in proximity to the object of interest. Gosselin *et al.* [25] provide evidence on the ability of blind people to use proprioception to guide hand orientation and to make rapid corrections of hand orientation during movements.

Hand-guided object localization.

Motivated by the observation above and prior literature in proprioception, we analyze existing datasets from blind users in the context of object recognition to examine the presence of the user's hand in photos. Moreover, we investigate the effectiveness of a hand-guided object localization approach for object recognition, where an object is localized based on its proximity to the hand, cropped and forwarded to the object recognition model. By using the hand as a guide to the object of interest, we not only ensure that the object is included in the frame, but we also reduce the effect that the background might have on the recognition model. This is natural since the object is cropped and the background is removed from the image, as shown in Fig. 1. To localize the object of interest based on its proximity to the user's hand, we use convolutional neural networks (CNNs) tofirst train a hand segmentation model and then fine-tune it to learn to locate the center of the object in

proximity to the segmented hand. The image is then cropped to include the located object of interest. We explore the effectiveness of this approach in the context of teachable machines, where the quality of the image impacts accuracy not only during the prediction step but also in training [36]. Specifically, we show that this approach could improve the accuracy of teachable object recognizers in real-world environments where objects lay in a cluttered background. With an extensive error analysis, we provide deeper insights into the feasibility and challenges of using existing egocentric datasets from the sighted population. Tofurther research in this direction, we make our dataset as well as the output of our models for the error analysis publicly available.

2 RELATED WORK

Object recognition and camera manipulation for visually impaired people have been extensively studied for accessibility. We discuss prior work that we draw upon to both inform our analysis and contextualize the implications of our results. Moreover, we present state-of-the-art deep learning approaches, which guide our model, from hand recognition in egocentric vision, a sub-field of computer vision focusing on the analysis of images and videos typically captured by a wearable camera.

Object Recognition for the Blind

Researchers have explored assistive technologies that use camera input to help visually impaired people identify surrounding objects. The following surveys prior work to identify the characteristics of these solutions and to understand: the diversity of user input, the mechanisms underlying prediction, and sources of training examples for deployed computer vision models. Table 1 presents representative examples from 2010–2018 focusing mainly on real-world applications, though similar patterns may be found when examining larger surveys [33, 40, 46]. While a few examples are similar to barcode scanners, requiring adhesive tags to be attached to objects, (*e.g.*, [17, 68]), the general trend is to use camera stream as an input for prediction with a single photo [4–6, 10, 15, 18, 19, 24, 65] or real-time video [5, 7, 9, 15, 53, 65, 67].

Prediction using computer vision models is common practice (Table 1). However, current limitations in object recognition make it impossible to build a “super” object classifier to recognize all possible object instances of interest for all blind people. Thus, such solutions are typically restricted to a few object instances (*e.g.*, US currency [18]), adhesive tags (*e.g.*, [17, 68]), objects whose images are available on the web or a database (*e.g.*, [24, 73]), and generic object classes (*e.g.*, [5, 15, 67]). To overcome some of these limitations, applications are either backed up (*e.g.*, [65]) or fully supported (*e.g.*, [7, 9–11]) by human respondents.

Most recently, researchers are exploring teachable object recognition which allows users to personalize a recognizer with their objects of interest. Early approaches required training examples provided by sighted users [63]. Subsequently, models trained by users with visual impairments have shown promise given well-framed training examples [4, 36, 62]. As there are few datasets with photos of objects taken by blind users, most models are trained on photos taken by sighted people. Such photos may exhibit different background clutter, scale, viewpoints, occlusion, and image quality than those taken by blind users at prediction time.

Informed by this design space, we explore the presence of users' hands in photos during **prediction** for a crowd-sourced solution, such as VizWiz [11], and a teachable solution, such as Sosa-Garcia and Odone [62]; they released their datasets (Sec. 4). Moreover, we explore the effectiveness of our hand-guided object localization in the context of **training**, by replicating prior work from Kacorri *et al.* [36] on a new small but rich benchmark dataset (Sec. 5).

Camera Manipulation for the Blind

We consider how challenges in camera manipulation faced by blind people have been studied in prior work. While not restricted to object recognition, prior studies analyzing photos taken by blind people have reported on difficulties related to blurriness, lighting, framing, composition, and overall photo quality [1, 12]. To overcome some of these challenges, researchers have leveraged models of visual attention [32] to either provide feedback to the user for better framing [66] or to select a high-quality photofrom a video stream [73]. Others have explored solutions that can detect blurriness and inform users about camera tilt [28]. While rarely discussed in the context of object recognition, such approaches assume that the users can spatially localize an object and aim the camera appropriately. Thus, they have limited utility in the case of partially included or out-of-frame objects of interest, cluttered backgrounds, and the presence of multiple objects.

Some reported strategies that blind users employ for taking good photos [1, 2, 28, 34] include: (i) positioning the camera to the center of an object and slowly pulling away while trying to keep it in frame, (ii) making an educated guess on where to point the camera, and (iii) taking several shots, hoping some will include the object. Jayant *et al.* [34] built upon the first strategy to guide users by asking them to take a first photo of the object up-close and then tofollow verbal feedback to move the camera away. Although this method is an interesting alternative approach that can be used in our work, it is verbose and has not been evaluated in a real-world setup and the presence of multiple objects.

Even though the presence of users' hands in the camera frame (mentioned in Kacorri *et al.* [36]) is not explicitly reported above, we associate it with the second strategy: users leverage proprioception [25] to make an educated guess. Prior work on blind users' fingertip detection to estimate the area of interest in a camera frame for inaccessible physical interfaces [26] or color detection [47] provides additional evidence for the potential of our approach.

Hands in Egocentric Vision

Egocentric vision analyzes images providing a first-person view from a wearable camera typically mounted on the user's body or, as in our context, a mobile camera hand-held by a blind user. As users' hands in egocentric vision can provide context about intentions, actions, and areas of interest, hand recognition is an active area of research in the computer vision community (*e.g.*, [13, 30, 72]) with many applications in virtual reality (*e.g.*, [69, 70]), augmented reality (*e.g.*, [29, 45, 54]), as well as within our community [16, 26, 47, 49, 52]. To our knowledge, this is the first study exploring hand detection in the context of object recognition for the blind.

Our hand-guided object localization leverages prior work by Ma *et al.* [44], which uses a twin-stream CNN to detect egocentric activities. Specifically, we adopt one of the network streams for our recognition pipeline (Fig. 1). However, we opt for a different network architecture using FCN-8s [42] to improve our model's understanding of hands in the first-person view. Using photos from sighted users with a head-mounted camera, we explore its potential for photos taken by both blind and sighted users using a mobile camera over a larger number of object instances.

3 A HAND-GUIDED OBJECT RECOGNIZER

We introduce a hand-guided object recognizer comprising three deep learning models (Fig. 2): hand segmentation, object localization, and object classification. Our architecture is based on prior work on egocentric activity recognition [44], though we use a high-precision fully convolutional network, FCN-8s, proposed by Long *et al.* [42].

Typically, thousands of annotated images are required to train such models. Therefore, in training our system, we make use of any existing pertinent datasets (Sec. 4) and create a new one for richer insights and error analysis (Sec. 5).

Hand Segmentation Model

Our hand-guided object recognizer is based on the intuition that objects of interest will appear in the vicinity of the user's hand. Thus, to estimate the center of the object of interest, we need to locate the hand within the frame. We train a hand segmentation model to perform this task.

Specifically, we use a fully convolutional network (FCN-8s), shown to perform well for fine-grained segmentation [42]. High performance is crucial to our approach as it requires highly accurate information on hand shape and poses. As shown in Fig. 3(a), our model is trained with images annotated with two labels, *background* and *hand*, where we estimate the class of each pixel; that is, whether the pixel belongs to the background or the hand. The model is optimized using the cross-entropy loss function.

Object Localization Model

Employing transfer learning [43, 48], we build an object localization model by "fine-tuning" our hand segmentation model. Intuitively, lower layers of the segmentation model have learned to identify image pixels corresponding to a hand. Thus, these features can be re-purposed to estimate the center of the object relative to the hand pixels.

Specifically, we keep weights for the first five layers in the segmentation model, shown to learn hand related features [44], and re-train the rest. Training images are annotated with a Euclidean heatmap mark indicating the object's center (Fig. 3(b)). Heatmap annotation has been shown to be more robust than pinpointing location coordinates regarding model training [44, 50]. Using cross-entropy loss, we train the localization model to predict the class of each pixel in the final layer: the background or the center location[1].

Annotating with a heatmap, the model learns which pixel belongs to the *possible center locations*. Given an image, we use this model to localize the object center and crop the image based on the estimate (blue overlay in Fig. 3(b)).

Object Classification Models

Automatic object recognition spans across: (i) pre-trained models on a static number of object classes, which we call generic object recognizers (GOR); and (ii) personalized models with teachable machines, allowing end-users to specify objects of interest and provide a small number of training examples, which we call teachable object recognizers (TOR).

We explore feasibility of hand-guided object localization for object recognition in both contexts: *generic* and *teachable*. The former uses state-of-art models trained on large data. In the teachable case, models are trained on the fly by users; thus we expect the effect of hand-guided localization to be larger as image quality affects both training and prediction.

GOR: We use Google's Inception V3 [64] pre-trained on the 2012 ILSVRC dataset [58] with 1,000 object classes.

TOR: We use transfer learning to re-train[2] GOR for 19 object classes using our benchmark dataset described in Sec. 5.

4 EGOCENTRIC HAND-OBJECT DATASETS

Training and evaluating each component requires an abundance of data. Ideally, the data are egocentric images from diverse groups of visually impaired people with mobile cameras to identify real-world objects using both GORs and TORs. Moreover, images require detailed annotations such as hand masks, object center, and object name (Fig. 3(a) and Fig. 3(b)). Such a dataset is not currently available.

To mitigate this, we identify related datasets, extend them with additional annotations, and create a new benchmark dataset (see Sec. 5). Table 2 summarizes these datasets along the number of images, distinct objects, and people, the level of people's vision, camera perspective, and environmental setup (examples shown in Fig. 4). Environments are reported as "wild" for images captured in the real world, and "vanilla" for images in a uniform laboratory setting where objects tend to occur with a plain background.

GTEA.—Georgia Tech Egocentric Activity (GTEA and GTEA Gaze+) datasets [41] are collected with 4 and 26 sighted people, respectively, through hat-mounted cameras[3]. For benchmarking activity recognition, they include annotation of actions being performed — frame-level gaze tracking data included in Gaze+ as well. They also provide hand mask

[1]We train both models with 10, 000 training steps, using the Adam optimizer [38], 10–6 learning rate, 0.9 beta1, 0.999 beta2, and 0.9 Adam epsilon.
[2]We fine-tune pre-trained Inception using 4, 000 training steps and gradient descent with learning rate 10–2. We augment data by flipping left/right, randomly cropping 10% of the margin, and varying brightness by ±10%.
[3]Only data with available hand masks (6 people) were used in GTEA Gaze+.

annotations, since hand shape and poses are informative in object and activity recognition [21]. We manually annotated the object centers[4].

Intel Egocentric Vision Dataset.—This dataset is closer to our task as it was collected for recognition of everyday objects handled by a person [55]. Similar to GTEA, two sighted participants recorded their activities using a shoulder-mounted camera. However, the default per-pixel annotations include both hand(s) and the object of interest. As our system recognizes a hand and an object separately, we manually generated hand-only masks and object-center annotations for a subset of this dataset (only 177 images from one participant)[5] and use them to train our segmentation and localization models.

EgoHands.—This dataset provides first-person interactions among four sighted people; two people formed a group, and each of two wore a glass-mounted camera to record the activities [8]. We use the 3, 752 images with hand masks to train our hand segmentation model.

Glassense-Vision.—This dataset is close to both our task and our target population as it was collected with three low-vision participants for TORs [62]. Thus, we use it to evaluate our approach. A total of 71 object instances were grouped along seven categories[6] of three geometrical types (flat, boxes, and cylinders). However, their training data were collected in a uniform setting, with objects on a white surface with no hands present. In contrast, the testing data were acquired while participants held the object on their hand. Therefore, our analysis only uses the testing data that capture hand-object interactions.

VizWiz.—This visual question answering dataset [27] provides a rich set of real-world images taken by visually impaired people along with questions to the crowd. While the images are not limited to object recognition, the dataset provides a unique opportunity to investigate how hand presence in real-world settings would lead to including and indicating the object of interest in the camera frame. Thus, we use this dataset to evaluate our approach.

5 TEgO: A NEW BENCHMARK DATASET

To investigate the feasibility of hand-guided object recognition for blind people, we create a new benchmark dataset, called Teachable Egocentric Objects (TEgO)[7]. By controlling for factors such as users, environments, lighting, and object characteristics, this small but rich datas*et al*lows us to explore and uncover potential strengths and limitations in the context of object recognition for the blind, which is not feasible with the existing datasets described above.

People.—Two individuals collected data over five weeks resulting in 11, 930 images. As shown in Table 2, the small number of people is comparable to those reported for similar egocentric datasets in computer vision, while the number of images is large. The first

[4]Additional annotations available at https://iamlabumd.github.io/tego/.
[5]Additional annotations available at https://iamlabumd.github.io/tego/.
[6]The cereals category in the Glassense-Vision dataset was unavailable at the time.
[7]TEgO is available at https://iamlabumd.github.io/tego/

individual (B) was a blind undergraduate (no light perception) with little experience using a mobile phone camera. The second individual (S) was a sighted student experienced in machine learning. Data from S serve merely as *an upper baseline for discriminative power of the system given that many of the objects shared visual similarities*; that is, we treat the data from S as ideal data and use them as a point of comparison in our analysis.

Objects.—Informed by prior work on TORs [36, 62], object stimuli (total of 19) were carefully engineered to cover a large group of categories with diverse geometries and functions. Objects within a category were deliberately chosen to have a similar shape and appearance (Fig. 5). By making the recognition task more difficult, we are hoping to anticipate more challenges faced in real-world deployments.

Environment and Lighting.—We collected our dataset in both simulated real-world (wild*) and more traditional lab (vanilla) settings. As shown in Fig. 1 and Fig. 4, “wild”-photos were taken against a cluttered background including a bookshelf, a kitchen, and many other objects. Whereas “vanilla”-photos were taken against a wooden surface next to a white wall. Three different lighting conditions are considered for testing in each environment: (i) indoor lights on, (ii) surrounding light with indoor lights off, and (iii) fash on with indoor lights off. The data collection lasted for many days while the effect of natural light was mitigated.

Procedure.—TORs are evaluated in two phases: training and testing. In training, the user sequentially provides a small number of photos: about 30 examples per object under each environment and lighting condition. To simulate testing, an object is randomly given to the user for a single shot. The random object assignment minimizes learning effects [36]. We iterate it until we test each object five times, resulting in a total of 19 × 5 test images at a time.

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

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

Annotations.—Beyond object labels, all images in our dataset are manually annotated with hand masks (Fig. 3(a)) and object center heatmaps (Fig. 3(b)).

6 EXPLORATORY ANALYSIS AND RESULTS

As discussed in Sec. 3, our system comprises three models:

Hand segmentation: The model was trained primarily on images taken by sighted people with hand mask annotations (a total of 5, 707) from GTEA, GTEA Gaze+, Intel Egocentric Vision, and EgoHands. Additionally, similar to the selective hand annotation in [41], we included randomly selected images (about 7%) from TEgO during the training phase (224 and 654 from S and B, respectively). By including some images from our dataset, we increase the potential of the model to perform well on the unseen data from S and B. It decouples the performance of the segmentation and localization models from the recognition model, and thus allows us to evaluate the utility of hands in recognition. Only the Intel Egocentric Vision dataset (177 annotated images) includes hands with a similar complexion to that of B. The majority of the data from sighted participants include hands with a complexion close to that of S (5, 530 images). Thus, we do not anticipate that the imbalance across B and S data, will bias our segmentation model in favor of B.

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

Object recognition: A GOR is pre-trained on a million of images from ImageNet [58], and TORs are trained on images from TEgO; each TOR is trained with 30 examples per object (total 19 objects) from S or B in a given condition: hand presence and environment. For example, a TOR for B is trained on 30 images per object taken in the wild with the screen button while using one hand to interact with the object.

Analysis on Existing Datasets from Our User Group

To provide further evidence for this work, we investigate whether hands are used as a natural interface by people with visual impairments for including and indicating the objects in the camera frame. We examine existing images from this population in the context of object recognition such as the Glassense-Vision [62] (vanilla) and VizWiz [27] (wild) datasets. In addition, we run our segmentation and localization models over these previously unseen datasets, report the proportion of images with such characteristics, and offer qualitative insights on their limitations.

Glassense-Vision.—Through manual examination, we find about 44% of images taken by people with low vision in the Glassense-Vision included the individual's hand when testing the performance of their teachable object recognizers in a vanilla environment. However, our model estimated that only 16% of images include a hand, appropriately cropping the object in 89% of these. By visually inspecting failed cases, where a hand was not identified, we find images with different ratio and orientation from those typical in our training examples. It highlights the importance of diversity in hand-segmentation training data for this population. Positive and negative examples of model outputs are shown in Fig. 6(a). While this analysis would benefit from a more quantitative approach for object detection, it requires ground-truth annotations which are currently unavailable for the dataset.

VizWiz.—As discussed in Sec. 4, VizWiz provides a richer, more realistic object recognition scenario as blind or low vision real-world users post images with associated questions to the crowd. To gauge what portion of VizWiz data are object recognition tasks, we note that most questions begin with "what". Based on prior analysis on this dataset [27], we estimate that around 29% of questions are related to object identification: about 8, 700 of 31, 173 images. While it is impractical to manually inspect all 31, 173 for the presence of a user's hand, we run our object segmentation and localization models on the VizWiz data and detect a total of 1, 548 images that include a user's hand for the object of interest. While still an estimate, this result indicates that, in at least 18% of images regarding object identification, real-world end users use their hands as a natural interface to indicate the object of interest. Figure 6(b) depicts examples of our localization model's positive and negative results on VizWiz. By visually inspecting 6, 000 random samples not detected by our hand model, we identify only 36 images (<1%) where users' hands were present. The majority (24) includes just the fingertips, and one image includes a previously unseen hand shape. Some images were blurry (9), and some had the camera flash on (9) — a lighting condition not present in any of the training sets. When visible, objects' shapes were: relatively flat (*e.g.*, paper, currency, gift card, newspaper, DVD cover), roughly cylindrical (*e.g.*, pen, cigarette, cans, bottles, mug), and nearly rectangular (*e.g.*, smartphones, laptop, keyboard).

Analysis on Our New Benchmark Dataset TEgO

For our analysis, we first separate the S data from the B data in TEgO, then divide their data into three subsets based on the method applied to the training and testing images — cropped-object (*CO*), hand-object *(HO)*, and object (*O*). The HO and O methods include original images taken by S and B. *HO* contains images where S and B are holding the object, while *O* contains the rest. *CO* consists of HO images cropped by the object localization model; that is, the CO images are extracted from the HO images.

Model Performance.—We explore the potential of the hand-guided recognition approach in the context of teachable object recognizers by comparing recognition performance of TORs trained on the CO images to those trained on the original HO and O images, respectively. Each model is trained on around 30 images per object at a given condition (*e.g.*, environment, button), and its accuracy is calculated on five testing images per object (total of 95) in the same condition. The average accuracy across models in HO, CO, and O is reported in Fig. 7, with error bars denoting standard error across multiple sets of testing results for S and B — 12 sets for S and 24 sets for B (due to the button settings).

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

Effect of Environment.—Fig. 8 shows a breakdown of the models' performance across vanilla and wild environments. As expected, overall model performance is lower in the wild, where cluttered backgrounds tend to be present. In the vanilla environment, we observe that, in general, models trained on images where S and B hold the objects (HO and CO) perform better than those where they don't (O). However, the utility of our approach seems more pertinent to the wild environment, where CO achieves, on average, 67% accuracy for B and improvement of 12% and 14% over HO and O, respectively. These results highlight the potential of our approach, given that photos of objects taken by people with visual impairments in the real world tend to include cluttered environments, as illustrated with the VizWiz dataset (Sec. 4).

Effect of Sample Size.—We explore the potential of our approach for training with highly limited sample sizes of 1 and 5 with k-shot learning in Fig. 9. Similar to Kacorri *et al.* [36], we observe that, on average, the model performance increases with the sample size. More importantly, we note that CO tends to outperform HO and O consistently across sample sizes.

Teachable vs. Generic.—We explore whether the positive effect of the CO method over the HO and O methods, observed in teachable object recognizers (TORs), carries on to a generic object recognizer (GOR). As discussed in Sec. 3, we use as our GOR the Google's Inception V3 model [64]. Since a GOR is not trained on the labels of TEgO, its accuracy score should not be compared with those of TORs, directly. As shown in Fig. 10, to allow for the comparison, we use V-measure [57] by comparing desirable properties of the two, such as consistency of their predictions given images of the same object, and ability to distinguish between two different objects. Similar to Kacorri *et al.* [36], we observe that a TOR achieves higher V-scores than a GOR. This is not surprising since a TOR model is fine-tuned to the users, their environments, and the number of objects. However, we did not anticipate *CO* and *HO* underperforming *O* in the GOR case. We suspect that the presence of hands in the HO and CO images, as well as close-up cropped images in the CO, may have induced some confusion in the GOR.

Error Analysis.—Interested on how to improve our method, we focus on hand segmentation and localization errors (Fig. 11) as well as object features affecting recognition (Fig. 12).

Hands out of frame.: While S and B held objects in all CO images, the B's hand is not present in nine images; but, the objects are. These were taken in the wild environment and included relatively large objects such as soda bottles (6) and cereal boxes (3). The localization model correctly identifies object centers on seven of them and failed to localize any on the other two, which partially include the object.

Hand segmentation errors.: Among 855 testing images collected in the vanilla environment, hands are partially segmented on 11 images for S and 9 for B. Even with partial segmentation, the localization model successfully infers the object centers. In 18 images from S and 29 from B, the segmentation model misclassifies small non-hand parts of the image in addition to correctly segmenting the hand. The localization model successfully

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

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

Discriminative features of objects.: The stimuli objects in TEgO were engineered to be diverse in terms of shape and function while sharing similarities to make recognition challenging. We illustrate aggregated results on misclassified testing images as a confusion heatmap (Fig. 12). We observe that some of the highest confusions are among objects within the same category that are cylindrical or share visual features such as soup can and mandarin can, salt and oregano, cheetos and lays. It indicates room for better performance in the recognition models (TORs) independent of the quality of cropped photos fed by our localization model.

7 DISCUSSION

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

Implications

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

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

User interactions.—Our analysis on existing datasets from people with visual impairments in the context of image recognition provides further evidence on the utility of hands as a natural interface for including and indicating the object of interest in the camera

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

Representative datasets.—While there is a recent trend of data sharing in accessibility that we embrace (*e.g.*, [22, 31, 37]), privacy concerns often prevent researchers from sharing images taken by people with visual impairments outside the lab. This work illustrates the potentials and limitations of leveraging existing data from a non-representative population such as sighted people. Beyond a population mismatch, we faced the following challenges: bias towards lighter hand complexions; bias towards male individuals (13 male; 2 female); lack of wrinkles, tattoos, and jewerly in the hands; and limited annotations for new but related learning tasks. We contribute towards representative datasets by making our data, TEgO and other annotations, publicly available.

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

Limitations

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

Due to the lack of available hand data with darker complexions, we included more hand data from the blind individual in training our segmentation model. While this might have biased our hand segmentation model, we do not see that this effect carries over into the segmentation (more errors are observed on images from the blind individual). We suspect that additional data can also help improve the performance of our approach by allowing to encode diverse hand shapes in the localization better. Prior work on hand-object manipulation [14, 23, 39] argues that people tend to use same or similar hand shapes for a given object shape and function.

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

Due to the lack of hand annotations for the Glassense and VizWiz datasets and available tools for understanding model behaviors, part of our analysis was based on subjective visual inspection. For a more objective analysis, we are currently exploring the compatibility of available approaches, such as Grad-CAM [59] and LIME [56] for explainability.

8 CONCLUSIONS AND FUTURE WORK

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

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

ACKNOWLEDGMENTS

The authors would like to thank Ebrima Jarjue, Dan Yang, June Xu, and Simone Pimentofor helping with the dataset; Kris Kitani and Chieko Asakawa for valuable discussions; as well as Jonggi Hong and the anonymous reviewers for insightful comments on earlier drafts of this paper. This work is supported by NIDILRR (#90REGE0008).

REFERENCES

[1]. Adams Dustin, Kurniawan Sri, Herrera Cynthia, Kang Veronica, and Friedman Natalie. 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.

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

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

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

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

[6]. VocalEyes AI. 2017 Computer Vision for the blind. http://vocaleyes.ai

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

[8]. Bambach Sven, Lee Stefan, Crandall David J, and Yu Chen. 2015 Lend-ing a hand: Detecting hands and recognizing activities in complex egocentric interactions. In Proceedings of the IEEE International Con-ference on Computer Vision. 1949–1957.

[9]. BeMyEyes. 2015 Lend you eyes to the blind. http://www.bemyeyes.org

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

[11]. Bigham Jefrey P, Jayant Chandrika, Ji Hanjie, Little Greg, Miller Andrew, Miller Robert C,Miller Robin Tatarowicz Aubrey, White Brandyn, White Samual, et al. 2010 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.

[12]. Brady Erin L, Zhong Yu, Morris Meredith Ringel, and Bigham Jefrey P. 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.

[13]. Cai Minjie, Kitani Kris, and Sato Yoichi. 2018 Understanding hand-object manipulation by modeling the contextual relationship be-tween actions, grasp types and object attributes. arXiv preprint arXiv:1807.08254 (2018).

[14]. Cai Minjie, Kitani Kris M, and Sato Yoichi. 2016 Understanding Hand-Object Manipulation with Grasp Types and Object Attributes.. In Ro-botics: Science and Systems, Vol. 3.

[15]. CamFind. 2013 Search the physical world. http://camfndapp.com

[16]. Castellini Claudio, Tommasi Tatiana, Noceti Nicoletta, Odone Francesca, and Caputo Barbara. 2011 Using object afordances to im-prove object recognition. IEEE Transactions on Autonomous Mental Development 3, 3 (2011), 207–215.

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

[18]. EyeNote. 2010 Mobile device application to denominate Federal Re-serve Notes (U.S. paper currency) as an aid for the blind or visually impaired to increase accessibility. https://www.eyenote.gov

[19]. EyeSpy. 2015 The worldâĂŹs best object recognition mobile app. http://www.eyespy.com

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

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

[22]. Flores German H and Manduchi Roberto. 2018 WeAllWalk: An Anno-tated Dataset of Inertial Sensor Time Series from Blind Walkers. ACM Transactions on Accessible Computing (TACCESS) 11, 1 (2018), 4.

[23]. Gilster René, Hesse Constanze, and Deubel Heiner. 2012 Contact points during multidigit grasping of geometric objects. Experimental brain research 217, 1 (2012), 137–151. [PubMed: 22198529]

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

[25]. Gosselin-Kessiby Nadia, Kalaska John F, and Messier Julie. 2009 Evi-dence for a proprioception-based rapid on-line error correction mecha-nism for hand orientation during reaching movements in blind subjects. Journal of Neuroscience 29, 11 (2009), 3485–3496. [PubMed: 19295154]

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

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

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

[29]. Hollerer T and Lee T. 2007 Handy AR: Markerless Inspection of Augmented Reality Objects Using Fingertip Tracking. In 2007 11th IEEE International Symposium on Wearable Computers(ISWC), Vol. 00 1–8. 10.1109/ISWC.2007.4373785

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

[31]. Huenerfauth Matt and Kacorri Hernisa. 2014 Release of experimental stimuli and questions for evaluating facial expressions in animations of American Sign Language. In Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel, The 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland.

[32]. Itti Laurent and Koch Christof. 2001. Computational modelling of visual attention. Nature reviews neuroscience 2, 3 (2001), 194. [PubMed: 11256080]

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

[34]. Jayant Chandrika, Ji Hanjie, White Samuel, and Bigham Jefrey P. 2011 Supporting blind photography. In The proceedings of the 13th inter-national ACM SIGACCESS conference on Computers and accessibility. ACM, 203–210.

[35]. Kacorri Hernisa. 2017 Teachable machines for accessibility. ACM SIGACCESS Accessibility and Computing 119 (2017), 10–18.

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

[37]. Kacorri Hernisa, Mascetti Sergio, Gerino Andrea, Ahmetovic Dragan, Takagi Hironobu, and Asakawa Chieko. 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.

[38]. Kingma Diederik P and Ba Jimmy. 2014 Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[39]. Klatzky Roberta L, McCloskey Brian, Doherty Sally, Pellegrino James, and Smith Terence. 1987 Knowledge about hand shaping and knowl-edge about objects. Journal of motor Behavior 19, 2 (1987), 187–213. [PubMed: 14988058]

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

[41]. Li Yin, Ye Zhefan, and Rehg James M. 2015 Delving into egocentric actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 287–295.

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

[43]. Long Mingsheng, Cao Yue, Wang Jianmin, and Jordan Michael I. 2015 Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791 (2015).

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

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

[46]. Manduchi Roberto and Coughlan James. 2012 (Computer) vision without sight. Commun. ACM 55, 1 (2012), 96–104. [PubMed: 22815563]

[47]. Mascetti Sergio, Gerino Andrea, Bernareggi Cristian, D'Acquisto Silvia, Ducci Mattia, and Coughlan James M. 2017 JustPoint: Identifying Colors with a Natural User Interface. In Proceedings of the 19th Inter-national ACM SIGACCESS Conference on Computers and Accessibility. ACM, 329–330.

[48]. Patricia Novi and Caputo Barbara. 2014 Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In Proceedings of the Computer Vision and Pattern Recognition.

[49]. Pavlovic VI, Huang TS, and Sharma R. 1997 Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. IEEE Transactions on Pattern Analysis & Machine Intelligence 19 (07 1997), 677–695. 10.1109/34.598226

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

[51]. Proske Uwe and Gandevia Simon C. 2012 The proprioceptive senses: their roles in signaling body shape, body position and movement, and muscle force. Physiological reviews 92, 4 (2012), 1651–1697. [PubMed: 23073629]

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

[53]. Recognizer LookTel. 2012 Instantly recognize everyday objects. http://www.looktel.com/recognizer

[54]. Reifnger Stefan, FrankWallhof Markus Ablassmeier, Poitschke Tony, and Rigoll Gerhard. 2007 Static and dynamic hand-gesture recogni-tion for augmented reality applications. In International Conference on Human-Computer Interaction. Springer, 728–737.

[55]. Ren Xiaofeng and Gu Chunhui. 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.

[56]. Ribeiro Marco Tulio, Singh Sameer, and Guestrin Carlos. 2016 Why should i trust you?: Explaining the predictions of any classifer. In Proceedings of the 22nd ACM SIGKDD international conference on knowl-edge discovery and data mining. ACM, 1135–1144.

[57]. Rosenberg Andrew and Hirschberg Julia. 2007 V-measure: A con-ditional entropy-based external cluster evaluation measure. In Pro-ceedings of the 2007 joint conference on empirical methods in natu-ral language processing and computational natural language learning (EMNLP-CoNLL).

[58]. Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, Huang Zhiheng, Karpathy Andrej, Khosla Aditya, Bernstein Michael, et al. 2015 Imagenet large scale visual recogni-tion challenge. International Journal of Computer Vision 115, 3 (2015), 211–252.

[59]. Selvaraju Ramprasaath R, Cogswell Michael Das Abhishek, Vedantam Ramakr-ishna, Parikh Devi, Batra Dhruv, et al. 2017 Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Local-ization.. In ICCV. 618–626.

[60]. SHERRINGTON Charles S. 1907 On the proprio-ceptive system, espe-cially in its refex aspect. Brain 29, 4 (1907), 467–482.

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

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

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

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

[65]. TapTapSee. 2012 Mobile camera application designed specifcally for the blind and visually impaired iOS users. http://www.taptapseeapp.com

[66]. Vázquez Marynel and Steinfeld Aaron. 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.

[67]. Aipoly Vision. 2016 Sight for Blind & Visually Impaired. http://aipoly.com

[68]. WayAround. 2018 The smart assistant for people who are blind. https://www.wayaround.com

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

[70]. Xu Deyou. 2006 A neural network approach for hand gesture recog-nition in virtual reality driving training system of SPG. In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, Vol. 3 IEEE, 519–522.

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

[72]. Zhang Yifan, Cao Congqi, Cheng Jian, and Lu Hanqing. 2018 EgoGes-ture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition. IEEE Transactions on Multimedia 20, 5 (2018), 1038–1050.

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

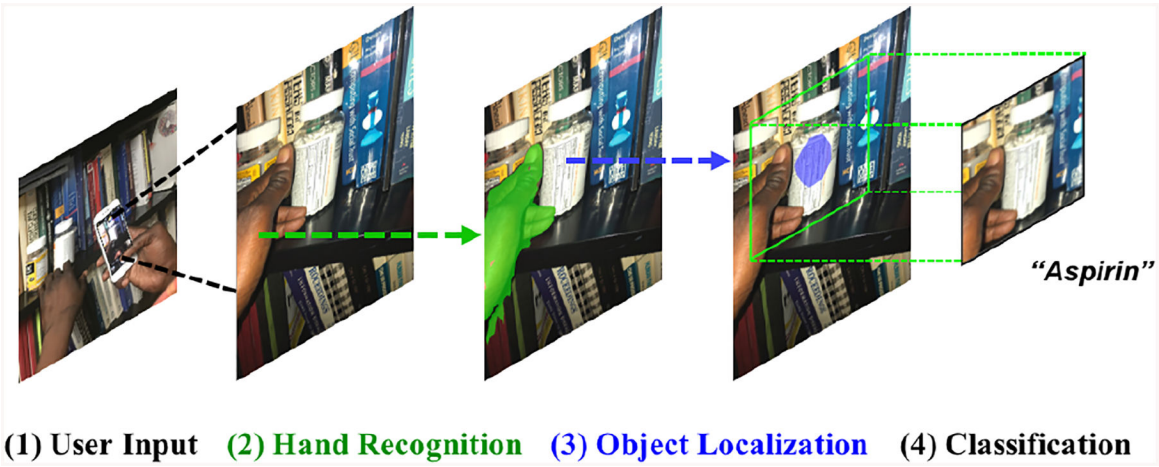

Figure 1:

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

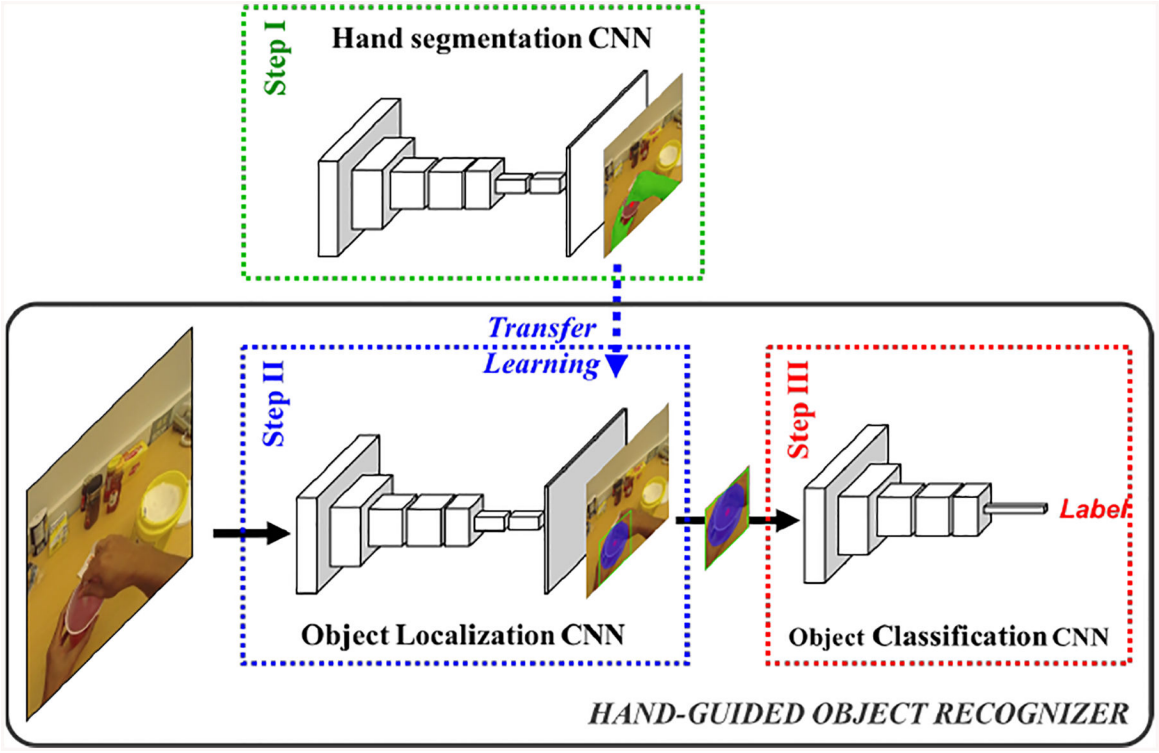

Figure 2:

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

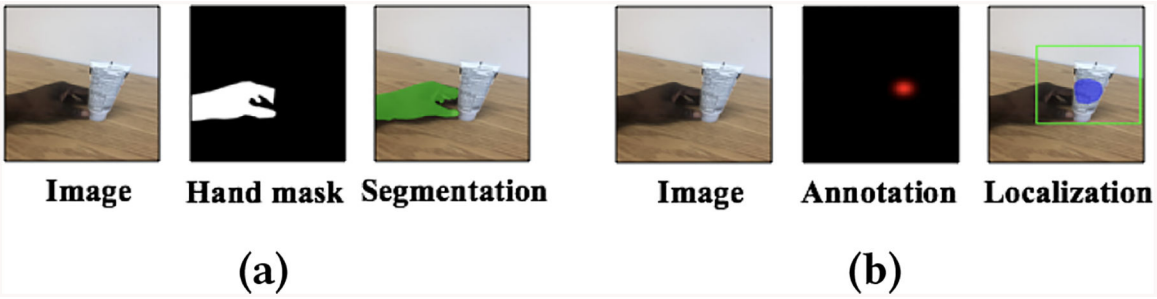

Figure 3:
An (input, annotation, output) example for our hand segmentation (a) and object localization (b) models.

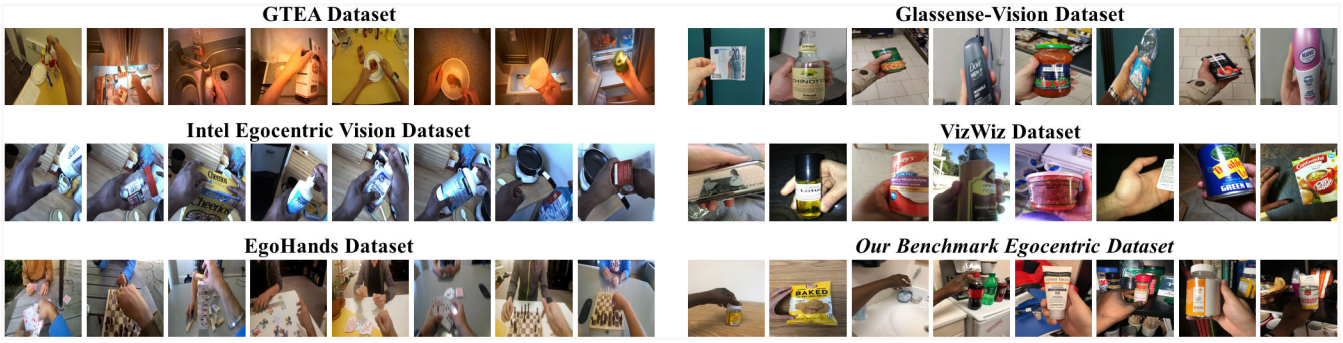

Figure 4:
Examples from each dataset. Glassense-Vision, VizWiz, and our benchmark examples are selected to include hands.

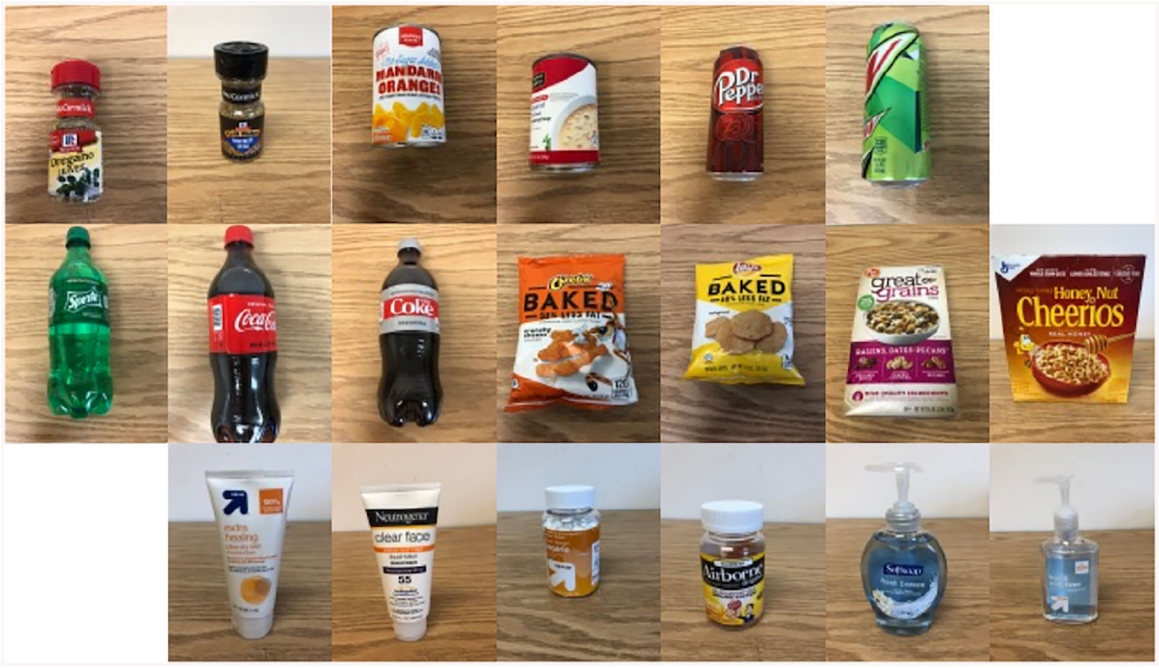

Figure 5:
Nineteen objects used in our data collection. Objects in the same category are displayed in proximity.

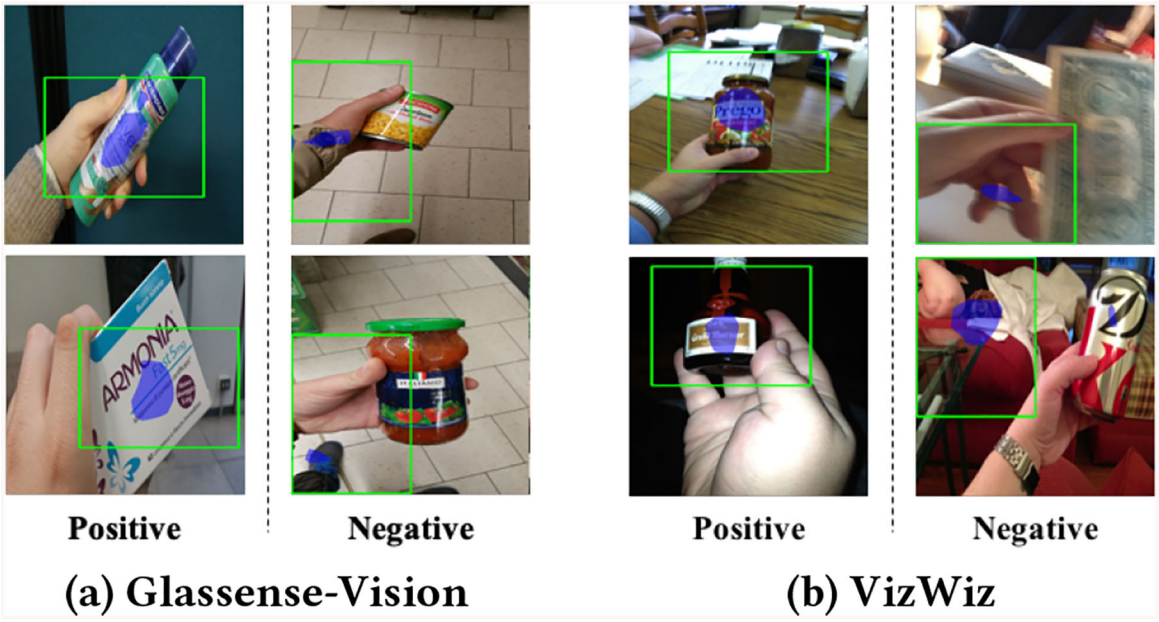

Figure 6:
Positive and negative outputs of our object localization model on the Glassense-Vision and VizWiz datasets.

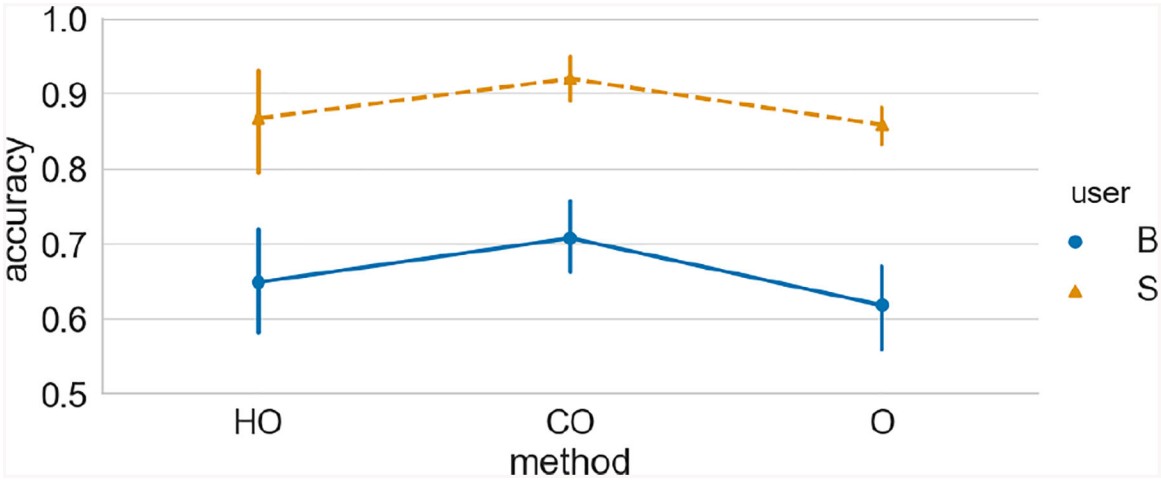

Figure 7:
Our hand-guided object recognition method (CO) tends to improve recognition accuracy on average for S and B compared to the original HO and O methods.

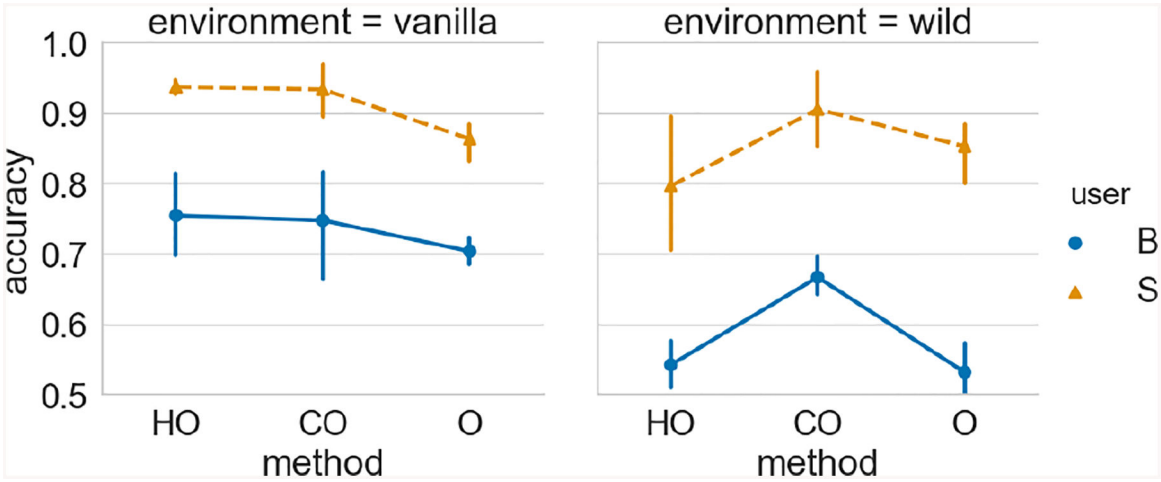

Figure 8:
Accuracy gain of our method (CO) over HO and O is more pertinent in cluttered backgrounds (wild).

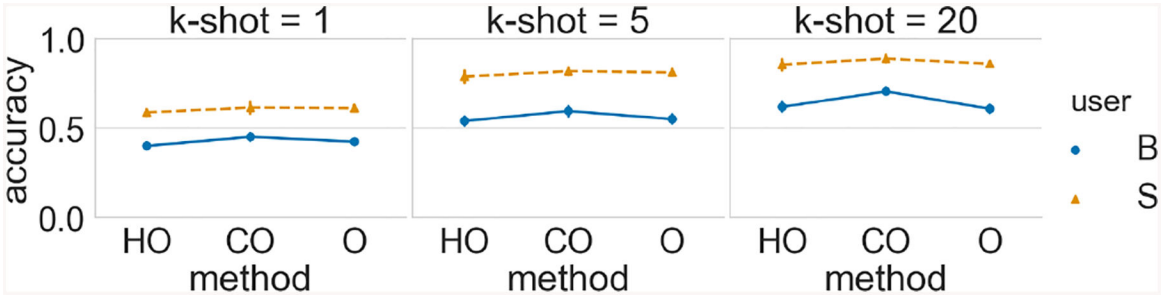

Figure 9:
On average CO outperforms HO and O consistently across training sample sizes $k = 1, 5, 20$.

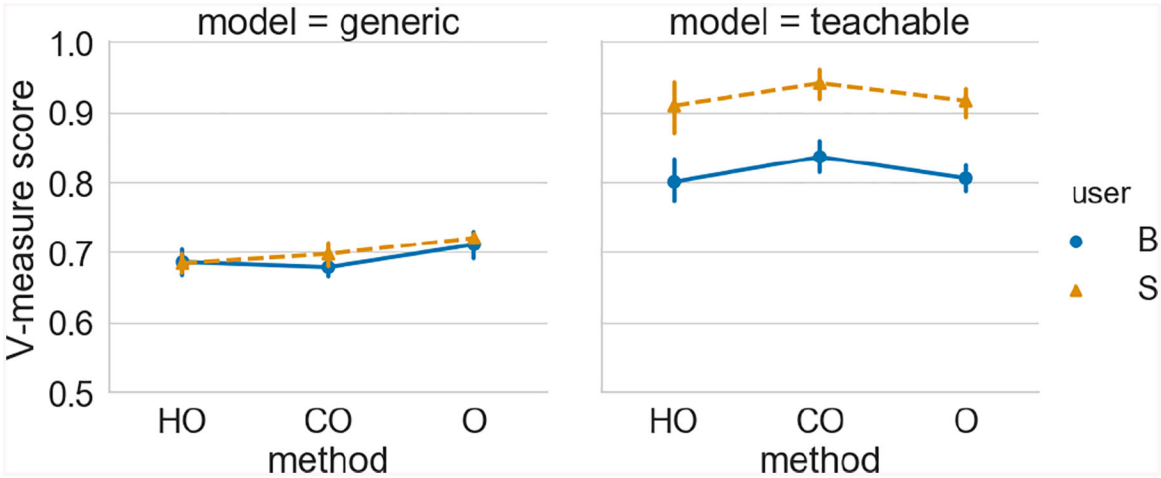

Figure 10:
Presence of hands (HO and CO) seems to have a different efect for generic vs. teachable models.

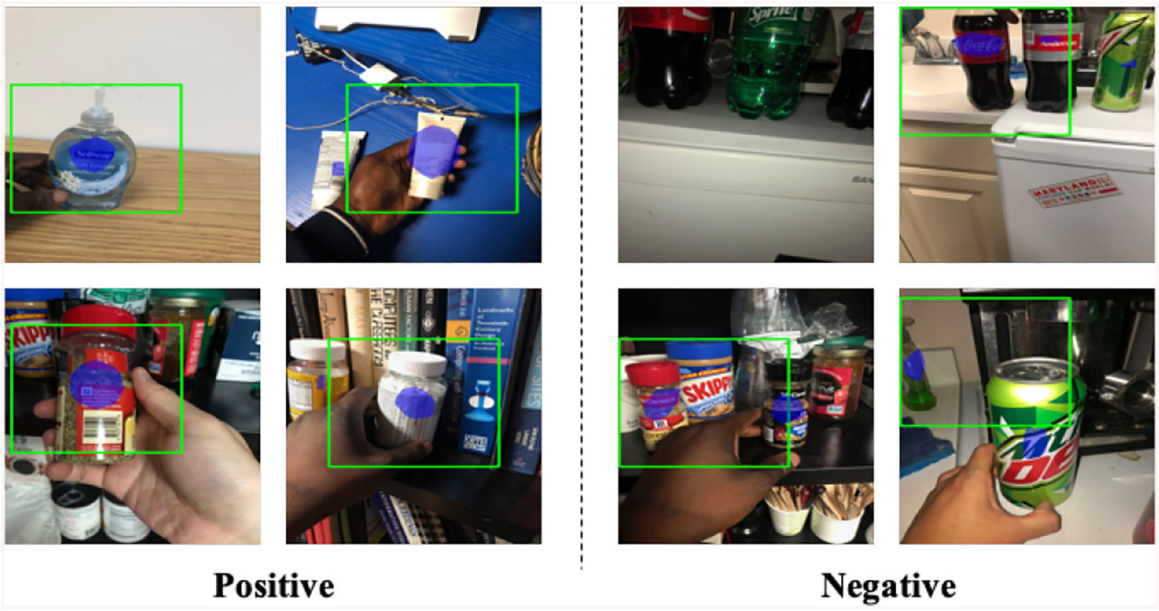

Figure 11:
Positive and negative results on TEgO, with outof-frame hands for some of the negative examples.

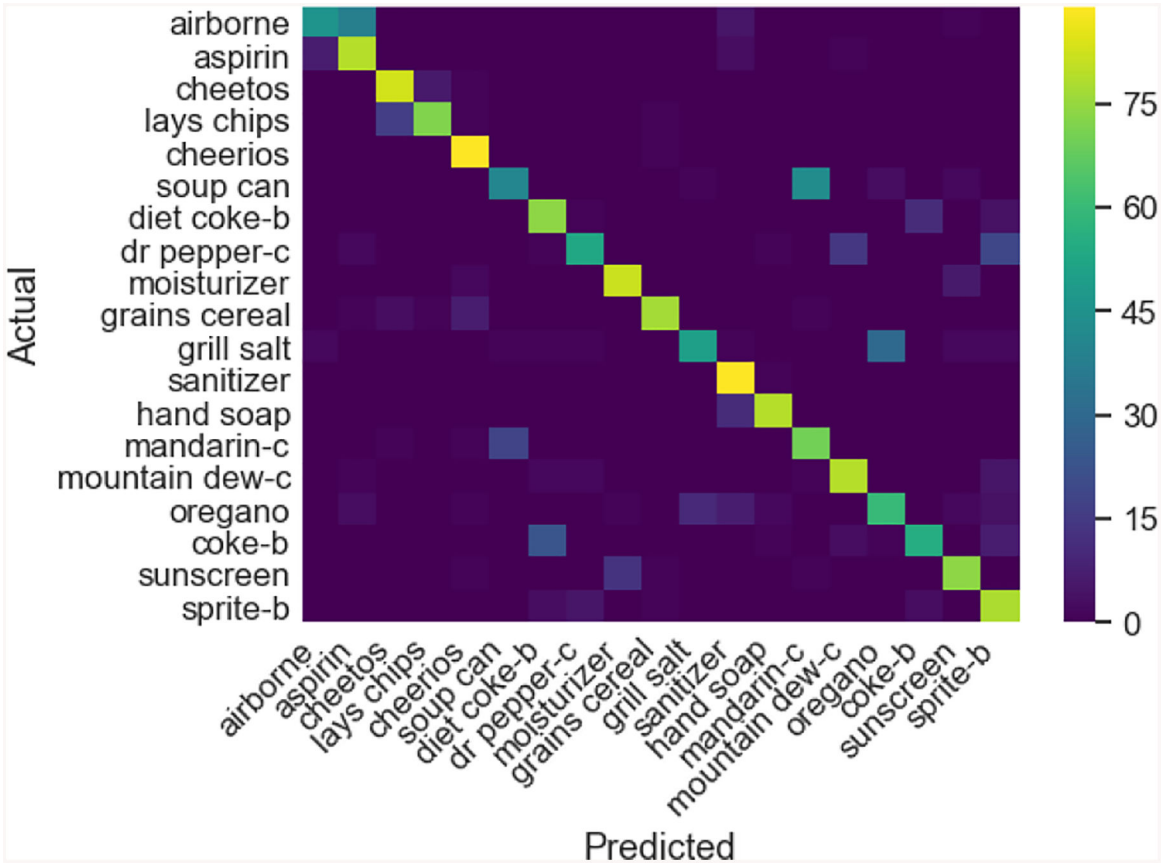

Figure 12:
Confusion matrix for the CO models showing that misclassifcation occurs within objects of similar shape. Cans and bottles are indicated as “-c” and “-b”, respectively.

Table 1:

Input, model, and prediction characteristics of prior work in object recognition for the blind.

System	Input		Trained Model			Prediction	
	img/video	tag	sighted	blind	personal	human	machine
VizWiz [11]	•					•	
Digit-Eyes [17]		•		•	•		•
EyeNote [18]	•		•				•
LookTel Rec. [53]	•				•		•
TapTapSee [65]	•		•			•	•
Zhong *et al.* [73]	•		•				•
Talking Goggles [24]	•		•				•
CamFind [15]	•		•				•
EyeSpy [19]	•		•				•
BeMyEyes [9]	•					•	
BeSpecular [10]	•					•	
Aipoly Vision [67]	•		•				•
Kacorri *et al.* [36]	•		•	•	•		•
Sosa-Garcia *et al.* [62]	•			•	•		•
Aira [7]	•					•	
Seeing AI [5]	•		•				•
VocalEyes AI [6]	•		•				•
WayAround [68]		•		•	•		•
Envision AI [4]	•		•	•	•		•

Table 2:

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

dataset	# images	# people	# objects	camera	vision	environment
GTEA [21]	663	4	16	on cap	sighted	wild*
GTEA GAZE+ [20]	1, 115	6	-	glasses	sighted	wild*
Intel Egocentric Vision [55]	177	1	42	on shoulder	sighted	wild*
EgoHands [8]	3, 752	4	-	glasses	sighted	wild
Glassense-Vision [62]	850	3	71	phone	low vision	vanilla
VizWiz [27]	31, 173	-	-	phone	blind and low vision	wild
Our benchmark	11, 930	2	19	*phone*	*sighted and blind*	*vanilla and wild**

Proc SIGCHI Conf Hum Factor Comput Syst. Author manuscript; available in PMC 2020 February 10.