ApplianceReader: A Wearable, Crowdsourced, Vision-based System to Make Appliances Accessible

Anhong Guo
HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
anhongg@cs.cmu.edu

Xiang 'Anthony' Chen
HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
xiangche@cs.cmu.edu

Jeffrey P. Bigham
HCI and LT Institutes
Carnegie Mellon University
Pittsburgh, PA 15213 USA
jbigham@cs.cmu.edu

CHI'15 Extended Abstracts, Apr 18-23, 2015, Seoul, Republic of Korea.
ACM 978-1-4503-3146-3/15/04.
http://dx.doi.org/10.1145/2702613.2732755

Abstract
Visually impaired people can struggle to use everyday appliances with inaccessible control panels. To address this problem, we present ApplianceReader - a system that combines a wearable point-of-view camera with on-demand crowdsourcing and computer vision to make appliance interfaces accessible. ApplianceReader sends photos of appliance interfaces that it has not seen previously to the crowd, who work in parallel to quickly label and describe elements of the interface. Computer vision techniques then track the user's finger pointing at the controls and read out the labels previously provided by the crowd. This enables visually impaired users to interactively explore and use appliances without asking the crowd repetitively. ApplianceReader broadly demonstrates the potential of hybrid approaches that combine human and machine intelligence to effectively realize intelligent, interactive access technology today.

Author Keywords
Non-visual interfaces; visually impaired users; accessibility; crowdsourcing; computer vision; wearable computers

ACM Classification Keywords
H.5.2 [Information interfaces and presentation]: User Interfaces - *Input devices and strategies*; K.4.2 [Computers and Society]: Social Issues - *Assistive technologies*

Introduction and Related Work

Visual impairment affects almost every activity of daily life, such as getting to where you want to go, reading information and recognizing objects around you, and participating in social interactions with other people [6]. Even simple and essential tasks like heating up food with the microwave oven is difficult when the functions of the buttons cannot be read and are not tactually differentiable.

To better understand the problem, we conducted an hour-long observation and an hour-long interview at the home of a visually impaired person. Our initial user research obtained several key insights that inform our system design for making appliances more accessible. *(i)* There is a strong need to make home appliances more accessible to visually impaired people, especially as touchpads are replacing physical buttons and taking away the tactile feedback, even though visually impaired people do not have problems locating the control area of the appliances. *(ii)* Visually impaired people often have to resort to social help, such as asking a neighbor or friend to help with using a home appliance: constantly seeking for help creates a social burden for both the help seekers and help providers, and often someone is not available to help when they are most needed. Thus it is important to find alternate solutions that can increase the independence of the visually impaired people in their home living. *(iii)* Labeling appliances with Braille seems a straightforward solution but means only environments that have been augmented are accessible, and may be unrealistic because fewer than 10 percent blind people in the United States can read Braille [1].

Two trends of research have sought to make home appliance more accessible to visually impaired people. First, using computer vision (CV), camera-based system such as the Access Lens 'reads' physical documents and lets a blind person listen to and interact with them [4]. CV in general excels at regularized problems, such as re-identifying objects seen before and recognizing regular black text on a white background. However, it struggles in many real-world, uncontrolled situations. This limitation was highlighted in the VizWiz project, which represents the second trend - crowdsourcing. Vizwiz allows a blind person to take a picture using a smartphone, speak to ask a question related to the picture, send both the picture and the question to the 'crowd' and get an answer in less than a minute [2]. Such a crowd-based system can answer a wide variety of questions that CV cannot handle, *e.g.*, does my outfit match, what are the buttons on this machine and how are they arranged on this device, etc. However, VizWiz's single question and answer is unwieldy for many tasks such as locating the right microwave oven buttons to heat up a meal. To achieve such tasks questions need to be asked and answered back and forth; and even with quick single responses, the accumulation of such dialogs across many everyday tasks could potentially slow down a user's daily routines. A few other systems have extended the types of questions or tasks Vizwiz can handle. For example, VizWiz::LocateIt [3] allows blind people to ask for assistance in finding a specific object, RegionSpeak [7] enables spatially exploring a region using a touchscreen. However, these two projects constantly rely on a handheld device rather than letting the user directly refer to or manipulate real world objects of interest.

In this project, we are developing a system that combines CV and crowdsourcing to help visually impaired people use home appliances' controls. When a visually impaired person encounters an appliance for the first time, he uses a smart eyewear camera, *e.g.*, the Google Glass, to capture

Figure 1: ApplianceReader is a wearable point-of-view camera-based system that combine crowdsourcing and computer vision to collaboratively label and recognize button controls on appliances, thus allowing visually impaired users to interactively explore and use appliances.

a picture of the device and then send it to the crowd. This picture then becomes a reference image: crowd workers label the overall layout of the appliance interface, and annotate the specific functions of each control. This is a one-time labeling and calibration step (Figure 2a). Later, when the visually impaired person wants to use the appliance, he simply turns on the Google Glass camera, looks at the control and hovers a finger over it. CV techniques use the initial crowd-labeled reference image to detect in real-time which control the user is pointing at and reads its functionality aurally (Figure 2b). With such instantaneous feedback (Figure 2d), the user can continuously search and locate the button he wishes to use by each time hovering or touching a button, listening to what it does, and deciding whether it is the right button to use. To handle uncertain situations, our system will also predict performance based on image quality, orientation and the resultant probabilistic distribution (Figure 2c), and decide whether to send the captured image to the crowd as an alternate back-up solution (Figure 2e). All the crowd-labeled images will be collected by the system and added to a library of reference images to increase the robustness of future recognition. Figure 1 demonstrates the usage scenario of ApplianceReader.

Our system design advocates a hybrid, reciprocal relationship between crowdsourcing and CV as they collaboratively tackle a real-world, real-time accessibility problem. First, our system summons the crowd to build up the basic knowledge of an arbitrary appliance, as this general task is beyond what CV can do alone. Based on this bit of crowd input, CV then handles and accelerates the regularized tasks of identifying which control the user is interacting with, avoiding the latency that would be required if asking the crowd repetitively. By bringing CV and the crowd's strengths and overcoming their weaknesses, this hybrid approach pursues an optimal use of the two technologies together that go beyond their individual capabilities alone.

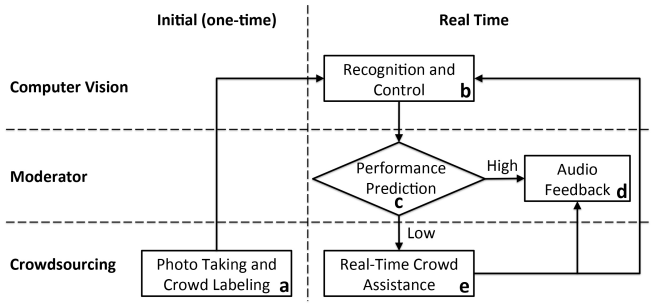

Figure 2: System overview, including initial and real-time stages using CV, crowdsourcing and moderator algorithms.

System Design and Implementation

To make home appliances more accessible to visually impaired users, ApplianceReader consists of *(i)* an initial labeling and calibration process to produce reference images of the appliance's controls, *(ii)* control recognition phases leveraging CV techniques on a combination of new camera input and the previously obtained crowd-labeled reference images to recognize and inform the user of the control he intends to use, and *(iii)* a performance prediction algorithm that coordinates CV and the crowd to complement one another, thus supporting fast and robust recognition in real-life situations. We have implemented the first two components, and have identified future steps for the complete implementation.

Initial Crowdsourced Appliance Labeling and Calibration
The first time a visually impaired user encounters an appliance, he uses ApplianceReader on his Google Glass to take a photo of the appliance (Figure 3a) and send the

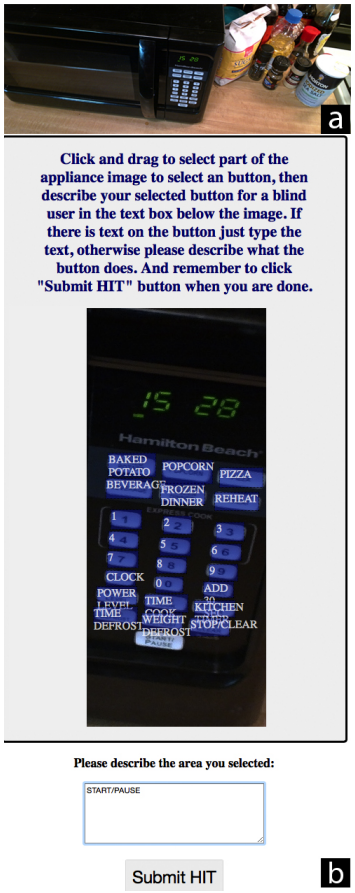

Figure 3: Initial (one-time) photo taking and crowd labeling of ApplianceReader: (a) an initial photo taken by the visually impaired user, (b) crowd workers verify the quality of the image and label interface elements.

image to a back-end server to be processed and pushed to the crowd for manual labeling.

In order to make the reference image most useful for CV algorithms, crowd workers are first instructed to draw a bounding box of the appliance interface, which will be cropped in the backend server to be used for later recognition. In this stage, the crowd workers are also asked to indicate the approximate number of controls, which will make it easier to distribute tasks and calculate compensation. Then, crowd workers are instructed to draw bounding boxes of the individual control elements (*e.g.*, buttons) within the interface area (Figure 3b); they will also type in text to describe what the control do (such as 'baked potato', 'start/pause'). We have built a PHP backend to handle uploaded images, the mechanism for the images to be sent to the crowd for labeling, and a database for labeled information storage. Our initial testing found that it takes approximately one minute to label one control. As an improvement to the crowd labeling process, ApplianceReader will collect answers from multiple workers in parallel. When there is disagreement, results will be aggregated using a majority vote in order to handle the uncertainty of the crowd.

One potential issue in this approach is that the quality of the reference image affects both crowd workers' performance and the accuracy of final results. This becomes a more serious problem considering the user is visually impaired and has difficulty taking a good photo of the object. To address this problem, we can, for example, let the user take continuous shots of the appliance and use stitching techniques to extract and combine candidate images in order to obtain better visibility of the controls [7]. With a larger coverage of the appliances' visuals, the system could potentially overcome the limitation of relying on one single shot of the object. In addition, we can ask the crowd to first assess the 'quality' of the image before performing labeling tasks, and prompt the user to take a replacement for low-quality images.

Recognition and Control Using CV + Crowd-Labeled Image
After the initial labeling and calibration, visually impaired users can now primarily rely on CV techniques with the crowdsourced reference image to 'see' and find specific controls. For example, to 'read' a particular button on an appliance, he simply turns on the Google Glass camera and hovers a finger over the button after locating the control area on his own. Using the reference image obtained earlier, ApplianceReader can identify which button the user is pointing at by firstly using SURF (Speeded-Up Robust Features) feature detector to figure out key points and feature vectors in both the reference (Figure 4a) and the input image (Figure 4b). Then the feature vectors are matched using FLANN (Fast Library for Approximate Nearest Neighbors) based matcher. By filtering matches and finding the perspective transformation between the two images using RANSAC (Random Sample Consensus), our system is able to localize the reference appliance image in the input image. In Figure 4b, the green bounding box is identified by transforming the corners of the reference image to corresponding points in the input image.

ApplianceReader then detects the fingertips location in the input image to be transformed to the reference image. One problem is that recognizing skin color using a fixed threshold is not robust because of changing lighting conditions. Therefore, ApplianceReader transforms the reference image to the plane of the input image using the perspective transformation (Figure 4d), then subtracts that with the image containing the user's finger

(Figure 4c), and obtains an image with mostly just the user's hand and finger left (they are the only expected difference between these images) (Figure 4e). After thresholding the image, ApplianceReader uses the largest contour of the convex hull to detect the fingertips location (Figure 4f) and to transform to the reference image (green line in Figure 4ab). This approach also reduces the size of the image to process to only the appliance interface, which reduces processing time.

Then by looking up the coordinates of the transformed fingertip location in the database of the reference image's labeled buttons, our system is able to detect which button is being pointed at (Figure 4, "2"). After identifying the control, the system converts its label to speech and announces its label. In initial testing with an input video stream at 640×360 resolution using a MacBook Pro, ApplianceReader can achieve a frame rate at 2 fps.

As next steps, if the image has low quality, *e.g.*, due to lighting conditions, orientation, occlusion, users are asked to retake the photo to improve it. If issues persist, a coordinating algorithm will take effect and bring in the crowd for help, as detailed below.

Coordinating and Combining CV and Crowdsourcing
On top of the aforementioned system components, we will also develop an algorithm that coordinates and combines CV and crowdsourcing based on various situations. Foremost, the algorithm will monitor and predict the performance of the CV techniques (Figure 2c); at times when the input images cause uncertain recognition results, it will provide the user with the option to 'ask the crowd' (Figure 2e). This approach will inevitably take a longer wait time but the returned crowd-labeled image can be added to the library of reference images and improve the robustness of the recognition. If a similar situation occurs in the future, this new reference image could be a close match and the answers can be directly obtained from its labels.

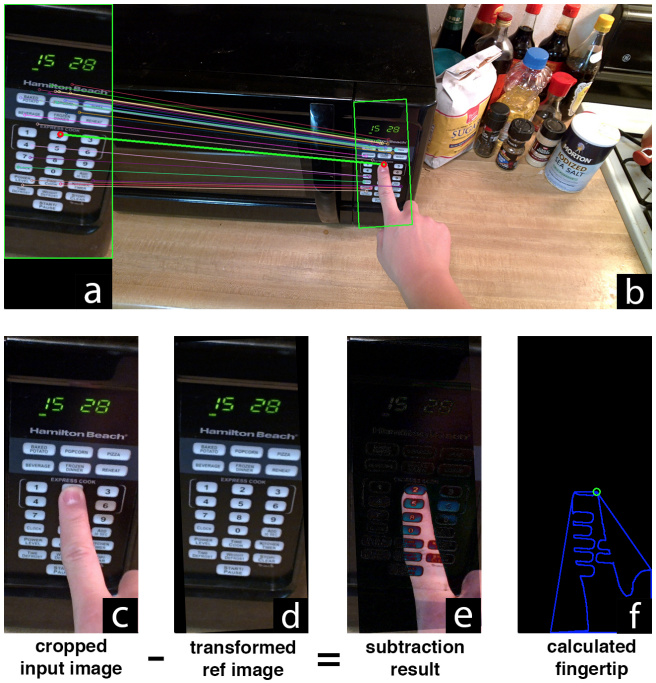

Figure 4: Real-time recognition and control using ApplianceReader. Recognition result, showing *(a)* reference image, and (b) input image. *(c)* Cropped input image. *(d)* Transformed reference image. *(e)* Subtraction result. *(f)* Calculated fingertip location.

Collectively, these reference images can also benefit a broader range of users when it comes to appliances in publicly shared space such as kitchens in office buildings. When a visually impaired user enters an unfamiliar office building and tries to use an appliance, he can simply

benefit from the reference images of someone who previously used the same appliance. When the images are geo-tagged, they can also help visually impaired users locate the nearest appliance they need to use.

Future Work

To date, we have implemented a Google Glass application that captures users' interaction with the appliances and communicates to a backend server, which then performs image uploading, labeling and the storage for the labeled information. We have also developed CV techniques to localize the appliances and identify the buttons a user is pointing at based on the reference image obtained earlier.

Our future work includes improving the backend crowd-labeling mechanisms, integrating and deploying the CV recognition algorithms to the server, and implementing the algorithms for coordinating and combining CV and crowdsourcing. We will also conduct a series of quantitative and qualitative evaluations. Since ApplianceReader is a combination of CV and crowdsourcing techniques, we will compare it with state-of-the-art systems using both technologies, such as VizWiz [2] and Chorus:View [5] on the crowdsourcing side, and OCR (Optical Character Recognition) on the CV side. The system will also be tested with visually impaired participants to evaluate its usability and elicit further design ideas and possibilities.

Conclusion

We are developing ApplianceReader - a wearable, crowdsourced, vision-based system to make appliances accessible for visually impaired users. ApplianceReader broadly demonstrates the potential of hybrid approaches that combine human and machine intelligence to effectively realize intelligent, interactive access technology.

Acknowledgements

This work was supported by National Science Foundation award #IIS-1149709.

References

[1] The braille literacy crisis in america. facing the truth, reversing the trend, empowering the blind. National Federation of the Blind, Jernigan Institute, March 2009.

[2] Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, ACM (2010), 333–342.

[3] Bigham, J. P., Jayant, C., Miller, A., White, B., and Yeh, T. Vizwiz:: Locateit-enabling blind people to locate objects in their environment. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, IEEE (2010), 65–72.

[4] Kane, S. K., Frey, B., and Wobbrock, J. O. Access lens: a gesture-based screen reader for real-world documents. In *Proc. CHI*, ACM (2013), 347–350.

[5] Lasecki, W. S., Thiha, P., Zhong, Y., Brady, E., and Bigham, J. P. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ACM (2013), 18.

[6] Manduchi, R., and Coughlan, J. (computer) vision without sight. *Communications of the ACM 55*, 1 (2012), 96–104.

[7] Zhong, Y., L. W. B. E., and Bigham, J. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proc. CHI*, ACM (2015).