FReAD: A Multimodal Interface for Audio Assisted Identification of Everyday Objects

Abhay Agarwal
Microsoft Research
No. 9, Lavelle Road
Bangalore, Karnataka, IN
t-abagar@microsoft.com

Sujeath Pareddy
Microsoft Research
No. 9, Lavelle Road
Bangalore, Karnataka, IN
t-supare@microsoft.com

Manohar Swaminathan
Microsoft Research
No. 9, Lavelle Road
Bangalore, Karnataka, IN
Manohar.Swaminathan@microsoft.com

CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
ACM 978-1-4503-4656-6/17/05.
http://dx.doi.org/10.1145/3027063.3053107

Abstract
Visually-impaired persons (VIPs) have employed tactility as a rich medium of interaction, exemplified by the ubiquity of braille and tactile boards. Yet, identifying everyday objects with tactility can still be challenging. We aim to augment a VIP's tactile faculties with modern computer vision, allowing users to touch and explore physical objects while receiving contextual, appropriate machine assistance. We present FReAD (Feature Reading Accessibility Device), a wearable device that provides audio assistance during tactile manipulations of everyday objects. The device can recognize the general taxonomy, printed text, and dominant colors of an object. Moreover, FReAD recognizes hand gestures, allowing objects to be classified automatically as soon as they are grasped. This information is then voiced to the user via a Bluetooth headset. FReAD employs a "belt-style" form factor that allows for hands-free interaction with objects and an unobstructed view of the user's frontal zone, while still encouraging seamless integration within the user's daily life.

Author Keywords
Wearable Computers; Accessibility; Human-Computer Interaction

Figure 1: (a) View of FReAD's belt-style wearable system, worn around the user's waist. The height is adjustable, but generally sits at the waist or belly. (b) View of the bluetooth headset. (c) View of user shopping while wearing the FReAD belt.

ACM Classification Keywords

H.5.2 [User Interfaces]: User-centered design; K.4.2 [Social Issues]: Assistive technologies for persons with disabilities

Introduction

Everyday objects are strikingly ambiguous to visually impaired persons (VIPs), given the nature of modern packaging and the lack of accessible tactile markings. To a VIP, a box of cookies may be indistinguishable from a box of toothpaste; a Granny Smith apple may be indistinguishable from a McIntosh apple. These problems are exacerbated by the industrial design of mass-manufactured goods, which place different goods in similar paper or plastic packaging.

In this work, we describe our efforts to augment a visually-impaired person with a computation system that may aid in discerning details of physical objects. We present FReAD, a belt-worn interface for VIPs. FReAD uses a combination of computer vision and crowd-sourced human labeling to provide short audio descriptions of objects automatically during the course of a user's normal tactile interactions. We term this near-instantaneous feedback system as "ambient" information. However, we also allow users to make verbal or gestural queries to request information on demand.

FReAD improves upon existing solutions for object recognition in several ways: it merges both computational and human-labeled information into a single interaction, it creates a simplified experience where existing user habits do not have to change significantly, and it addresses different criticisms of existing form-factors.

The primary components include a shallow depth-of-field (DOF) camera, a hand tracker (Leap Motion) and a wireless single-ear bluetooth headset. While our prototype runs on a backpack mounted laptop, the design can be easily modified to run on an embedded system.

Object Recognition for Accessibility

Specialized wearable hardware for object recognition has become available recently, such as Seeing AI [8] and OrCam [7]. These products, which both follow the Head-Mounted Display (HMD) form-factor, ostensibly free the user to engage with both their hands. In previous work [18], we evaluate a similar HMD system, but we identify problems with a head-centric model. Intuitively, fully blind individuals do not tend to point their heads toward the object of focus, limiting the capabilities of such systems. Individuals with limited sight fare better with HMDs, but in our work we preferred to design more general systems.

Smartphone apps exist for object recognition as well, several of which enjoy widespread use. These apps employ a variety of techniques, both computational and human-based, to create a spoken description of any photo taken by the user's phone. Apps such as Google Goggles [3] use computational approaches to classify images, provide a short textual description, and surface related internet content. Alternatively, apps such as VizWiz [11], CamFind [1], and TapTapSee [9] use on-demand human labelers to create a 3 to 10 word phrase per image. This approach addresses many of the drawbacks of purely computational systems, but increases the latency to between 10 and 20 seconds. This latency can be frustrating to users and possibly discourage continued use of the system. Moreover, VIPs often have trouble taking usable photos due to issues with angle, lighting, and focus.

Exploration of User Behaviors

We explored VIP behaviors to better understand how they make sense of everyday objects, and to qualitatively understand their usage patterns of mobile applications for object detection. Instead of conducting a large scale quantitative study as well, we refer to Brady, et. al [12]. We part-

Figure 2: (a) VIP using TapTapSee, a mobile app for object detection. He uses his right hand as a "ruler" to ensure the correct distance between the phone and the object. (b) Close-up shot of VIP using FReAD to identify objects in a controlled testing environment. The device naturally allows the user to explore the objects with their hands without interrupting the detection.

nered with two local schools for the blind for our study: Samarthanam Trust for the Disabled[1] and EnAble India[2]. We conducted informal studies of seven students, whose impairment ranged from partially to fully blind (late as well as congenital).

We provided subjects with a set of five objects to identify (a bag of potato chips, a water bottle, a soft drink can, a shirt, and a packaged razor blade). The purpose of the experiment was to observe how VIPs interact with objects to identify them.

We asked users to describe in as much detail as they could what the objects were, including brands and varieties, if possible. We gave half the subjects access to TapTapSee, an app for human-labeling that has been explicitly designed for the visually impaired. If the subject was unfamiliar with TapTapSee, we trained the subject until they could successfully identify an object with the app (5 to 10 minutes per subject). For packaged foods, we did not allow the subject to open the packaging during the study since subjects indicated to us that smell is an important aid for identifying food.

Without the aid of technology, subjects often mis-identified objects, due to the lack of any uniquely identifiable tactile features. Interestingly, when probed about their reasoning, the subjects would also display high degrees of certainty in their erroneous identification.

When provided with TapTapSee, users were usually able to correctly identify objects after 2-5 attempts (corresponding to between 30 seconds and 2 minutes). In attempts to overcome this frustrating inaccuracy, users displayed a wide variance in techniques for taking photographs. Some arranged the object specifically or used their hands to gauge focal distance (see Figure 2a). We also noticed a strong tendency toward tactile information-gathering in VIPs despite access to TapTapSee. In other words, the phone did not obviate the desire to touch and "confirm" the nature of the object.

These observations helped direct our focus to designing a wearable augmentation that could merge tactile exploration with computer vision approaches. While mobile apps do not explicitly prohibit tactile interaction, the interface discourages it by forcing the user to focus their hands on the process of photo-taking.

Implementation of FReAD

Form Factor

FReAD's device enclosure is shown in Figure 1. A belt buckle form-factor was chosen for its similarity to existing products, ability to withstand shakes and jerks, and its generally clear view of the user's grasping region. An example view from the camera as well as a resulting feature detection is shown in Figure 3. Furthermore, the form factor allows the device to be always on to minimize unnoticeable human error.

Functional belts or waist-worn pouches are pervasive with VIPs. For example, EnAble India provides all their students with a velcro waist strap to hold their cane. This form-factor improves on the head-mounted design since the user's do not need to use gaze to orient the camera. We considered cases in which this waist-worn design may be sub-optimal. If the user is seated at a desk, the camera may be below the table. However, the user may re-position the device to sit higher on their chest if they choose. If the object is sufficiently large, then the waist-mounted camera will not be

[1] http://www.samarthanam.org/
[2] http://enable-india.org/

Figure 3: Example text-recognition on object. (a) Raw frame captured(b) Approximate finger and palm positions are calculated, with correspondence mapping to camera frame (in yellow). (c) ROI estimation and cropping (in cyan). (d) Extremal-Regions classification on the ROI (in magenta). The final detection in the image is "NIVEA MEN CREME FACE BODY HANDS".

able to detect the hands as they will be outstretched or otherwise obscured. In these cases, the user is able to use verbal commands to request assistance from the device. In some cases, the user may not have the physical mobility to to buckle and unbuckle the device, but this problem applies to hand-held devices like a smartphone camera as well. Further research on this design is needed to understand the impact of fashion preferences, gender, and stigmas associated with outwardly obvious accessibility technology.

Components

FReAD combines a stereo-IR sensor for gesture detection (Leap Motion [5]) and a webcam with fixed, shallow DOF (Microsoft LifeCam Studio[3]). The DOF places the user's frontal region in focus while naturally blurring far-away objects, making later computation easier. The entire device and 3D-printed enclosure (without the laptop) costs around 14,250 INR ($212), but we estimate that our system, including an embedded processing unit, could be designed with inexpensive parts for as little as 10,000 INR ($150) Several mobile devices such as Project Tango [16] now have stereo image sensors that can perform spatial gesture recognition, making it conceivable to develop the application entirely on a pre-built device (as long as a suitable strapped enclosure is made).

Our software is implemented as an asynchronous multi-threaded application in Python using OpenCV [13]. Gestures, speech, and camera frames are received from our sensors triggering cascading dependent tasks in parallel. A schematic of the FReAD processing pipeline is shown in Figure 4.

[3] https://www.microsoft.com/accessories/en-us/products/webcams/lifecam-studio/q2f-00013

Computer Vision

We use Microsoft's Computer Vision Cognitive Services [6], which takes an arbitrary image and returns metadata such as objects detected, text detected, dominant colors, and a caption in natural language. Google's Cloud Vision API [10] provides similar functionality through its own separate implementation. The latency of requests made was fairly low, with a median latency of 3.2 seconds. This was observed to stem mostly from the method of internet connection because traces from the server reported the median latency of computation as 100 ms.

Crowd-Labeling

We use CloudSight [2] for human-aided captioning of camera frames, which returns a short phrase written by a human worker in a non-trivial amount of time. It may alternatively return a rejection, if the image is not suitable for captioning. From over 500 images collected by tests on VIPs, Cloudsight returned descriptions of minimum 1 and maximum 9 words. The median latency for a call was 12.73 seconds. The rejection rate was 7.6%, in which 5.7% were too close, 1.1% too blurry, 0.4% were too bright, and 0.4% were too dark. From an informal evaluation against TapTapSee, we found the rate of rejection of FReAD to be much lower. A longer study will be necessary to compare the overall correctness of the returned descriptions. However, the ease of positioning the object in front of the waist camera significantly lowered the burden of taking photographs.

The significant latency difference between computational and human-powered labeling may not be entirely detrimental in practice, since a user can rotate or re-align the object for multiple OCR calls in the time it takes to return a single crowd-label. Thus, by the time the crowd labeling arrives, the user often has a reasonable idea of some specific textual and tactile properties of the object in order

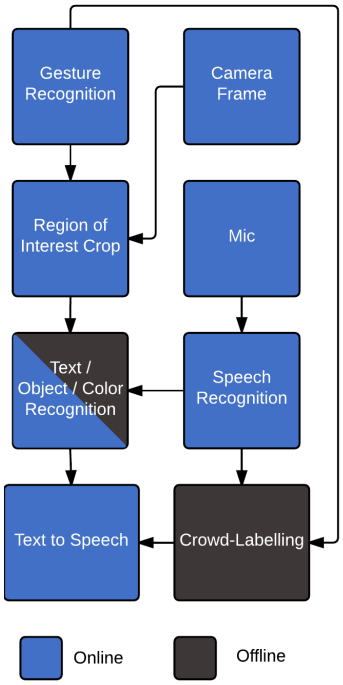

Figure 4: Architecture of FReAD

Figure 5: Programmed gestural metaphors. (a) A generic grasp yields a computational object detection that includes text and color. (b) A thumbs-up gesture after the object has been grasped triggers a crowd-labeling request.

to more clearly interpret the general description given by CloudSight. This stands in stark contrast to mobile applications, which provide no information between human-labeled responses.

All images are highly compressed prior to network transfer, which yields an image size between 20 and 100 KB. However, since sending a cloud call every frame is network bandwidth prohibitive, we use several computational techniques to prioritize on-board detection as much as possible. FReAD does not automatically send any frames to CloudSight due to the significant latency of responses, and is only triggered via an intentional interaction. For text recognition, we minimize cloud calls by predicting the frame's likelihood to contain text. To do this, the Leap Motion retrieves a point cloud of finger-bone positions in real space, which can easily be transformed into the camera's frame of reference. This allows us to quickly crop unwanted regions of the image. From the regions of the image likely to contain objects, we run a fast Extremal-Regions Text Detection classifier [17] that returns regions likely to contain text. If enough regions contain text, the frame is sent to the cloud for OCR processing. A pictoral representation of this process is shown in Figure 3.

Descriptions are verbalized through a discrete headset. Other devices have used bone-conducting headphones for a less invasive solution [15], however in our controlled environment this was not a significant factor. All our text is verbalized through an Indian English voice engine, which we found to be more comprehensible to Indian listeners than American English. However, it was observed that VIPs with higher computer literacy often preferred an American English voice, since it more closely matched the accent of their computer or phone.

Query form(s)	**Function**
"What color is this?" "Tell me what color this is."	Color
"Is there any writing?" "What's written here?" "Is there any text?"	Text
"What's in my hand?" "What am I holding?"	Object

Table 1: Programmed Verbal Queries. We adopt a broader entity-based model for query recognition that allows for multiple formulations of a certain intent.

Gestural and Verbal Interaction

FReAD is designed to provide information both ambiently (without an explicit request) and on-demand. It interprets the user grasping an object at chest level as a trigger for audio assistance. The system locally recognizes printed text and voices it. In the situation where a user wants a detailed description of an object, they may verbally query by asking special commands such as "What's in my hand?" or "What am I holding?", which trigger a cloud call to CloudSight. Users may also ask for specific information, such as asking "Which color is this?" to receive a list of dominant colors in the grasped object. Repeat calls to text detection can be made through commands such as, "Is there any writing?". In the case where a user desires a non-verbal method of querying, we implemented a "thumbs-up" gesture to signal a query to the human-labeling service. The gesture was chosen because it can be performed easily with one hand, has no negative cultural connotation in India, is unlikely to be affected by orientation and self occlusion of the hand and can be accurately detected from LeapMotion data. This

functionality is summarized in Table 1 and Figure 5. While the grasp detection would provide any (non-human labeled) information automatically, we implemented verbal queries to enable the user to demand information in times where the hands may be hidden or otherwise busy. Both the gesture set and the voice commands need to be refined with further user testing.

Audio cues help users easily frame objects within the camera's field-of-view (FOV). A periodic beep is given to the user when their hand is in the camera frame, while a different beep is given when the hand leaves the camera frame. In this way, the user can quickly assess whether their grasp and framing is correct. Users often placed their empty hands in the FOV then moved their hands around to get a sense of the dimensions of the FOV.

Conclusion

We have described FReAD, a device for audio-assisted identification of everyday objects. FReAD has shown promise in early trials and has the potential to offer a more natural user experience for VIPs, and of provide a richer description of everyday objects by mixing both computational and human-generated descriptions.

In a survey work on computer vision for object detection, Jafri et. al. [14] enumerated five properties needed for an interface to be successful with VIPs: it must be wearable, portable, real-time, noise-tolerant, and cheap. FReAD successfully meets all five criteria despite being at the prototype stage.

A Computational Sense of Touch

We attempted to augment a user's sense of touch rather than replace their lost vision. Despite the allure of computational sight, we feel that a more realistic enhancement is a computational sense of touch that can augment users' natural tendencies. This approach applies across types of visual impairments, and may affect the development of technologies for the sighted as well.

Future Work

We would like to test FReAD against other widely available accessibility solutions for object detection, comparing its accuracy, descriptiveness, and overall user experience. We have not studied its capabilities in outdoor scenarios, or studied its extended use. For this, a longer-term "in-the-wild" study is needed, where users use FReAD in their daily lives.

The device itself could be redesigned with a cheap, compact form factor that can communicate with or integrate with a device such as a smartphone. As embedded systems become increasingly capable of performing complex machine-learning tasks, it may be possible to run a larger fraction of object detection and OCR locally.

Since India is among the countries with the lowest data bandwidth costs [4], solutions like FReAD may reach a significant number of VIPs. In the future, we aim to explore how this device can be adapted to enable employment.

REFERENCES

1. 2017. CamFind App - Visual Search & Image Recognition API. (2017). http://camfindapp.com
2. 2017. CloudSight - Image Recognition API & Visual Search Engine. (2017). https://www.cloudsightapi.com/
3. 2017. Google Goggles - Android Apps on Google Play. (2017). https://play.google.com/store/apps/details?id=com.google.android.apps.unveil

4. 2017. ICT Facts and Figures 2016. (2017). http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx
5. 2017. Leap Motion | Mac & PC Motion Controller for Games, Design, Virtual Reality & More. (2017). http://leapmotion.com
6. 2017. Microsoft Cognitive Services - Computer Vision API. (2017). https://www.microsoft.com/cognitive-services/en-us/computer-vision-api
7. 2017. OrCam - See For Yourself. (2017). http://www.orcam.com/
8. 2017. Seeing AI Project - Pivothead Wearable Imaging. (2017). http://www.pivothead.com/seeingai/
9. 2017. TapTapSee - Blind and Visually Impaired Assistive Technology. (2017). http://taptapseeapp.com
10. 2017. Vision API - Image Content Analysis | Google Cloud Platform. (2017). https://cloud.google.com/vision/
11. Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: Nearly Real-time Answers to Visual Questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 333–342. DOI: http://dx.doi.org/10.1145/1866029.1866080
12. Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual Challenges in the Everyday Lives of Blind People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2117–2126. DOI: http://dx.doi.org/10.1145/2470654.2481291
13. Itseez 2014. *The OpenCV Reference Manual* (2.4.9.0 ed.). Itseez.
14. Rabia Jafri, Syed Abid Ali, Hamid R. Arabnia, and Shameem Fatima. 2014. Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. *The Visual Computer* 30, 11 (2014), 1197–1222. DOI: http://dx.doi.org/10.1007/s00371-013-0886-1
15. Chang-Gul Kim and Byung-Seop Song. 2007. Design of a Wearable Walking-guide System for the Blind. In *Proceedings of the 1st International Convention on Rehabilitation Engineering & Assistive Technology: In Conjunction with 1st Tan Tock Seng Hospital Neurorehabilitation Meeting (i-CREATe '07)*. ACM, New York, NY, USA, 118–122. DOI: http://dx.doi.org/10.1145/1328491.1328523
16. Eitan Marder-Eppstein. 2016. Project Tango. In *ACM SIGGRAPH 2016 Real-Time Live! (SIGGRAPH '16)*. ACM, New York, NY, USA, Article 40, 1 pages. DOI: http://dx.doi.org/10.1145/2933540.2933550
17. L. Neumann and J. Matas. 2012. Real-time scene text localization and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. 3538–3545. DOI: http://dx.doi.org/10.1109/CVPR.2012.6248097
18. Sujeath Pareddy, Abhay Agarwal, and Manohar Swaminathan. 2016. KnowWhat: Mid Field Sensemaking for the Visually Impaired. In *Proceedings of the 2016 Symposium on Spatial User Interaction (SUI '16)*. ACM, New York, NY, USA, 191–191. DOI: http://dx.doi.org/10.1145/2983310.2989190