ABSTRACT

Title of Dissertation:	HANDSIGHT: A TOUCH-BASED WEARABLE SYSTEM TO INCREASE INFORMATION ACCESSIBILITY FOR PEOPLE WITH VISUAL IMPAIRMENTS
	Lee Stearns, Doctor of Philosophy, 2018
Dissertation directed by:	Professor Jon E. Froehlich Department of Computer Science

Many activities of daily living such as getting dressed, preparing food, wayfinding, or shopping rely heavily on visual information, and the inability to access that information can negatively impact the quality of life for people with vision impairments. While numerous researchers have explored solutions for assisting with visual tasks that can be performed at a distance, such as identifying landmarks for navigation or recognizing people and objects, few have attempted to provide access to nearby visual information through touch. Touch is a highly attuned means of acquiring tactile and spatial information, especially for people with vision impairments. By supporting touch-based access to information, we may help users to better understand how a surface appears (*e.g.*, document layout, clothing patterns), thereby improving the quality of life.

To address this gap in research, this dissertation explores methods to augment a visually impaired user's sense of touch with interactive, real-time computer vision to access information about the physical world. These explorations span three application areas: reading and exploring printed documents, controlling mobile devices, and identifying colors and visual textures. At the core of each application is a system called HandSight that uses wearable cameras and other sensors to detect touch events and identify surface content beneath the user's finger. To create HandSight, we designed and implemented the physical hardware, developed signal processing and computer vision algorithms, and designed real-time feedback that enables users to interpret visual or digital content. We involve visually impaired users throughout the design and development process, conducting several user studies to assess usability and robustness and to improve our prototype designs.

The contributions of this dissertation include: (i) developing and iteratively refining HandSight, a novel wearable system to assist visually impaired users in their daily lives; (ii) evaluating HandSight across a diverse set of tasks, and identifying tradeoffs of a finger-worn approach in terms of physical design, algorithmic complexity and robustness, and usability; and (iii) identifying broader design implications for future wearable systems and for the fields of accessibility, computer vision, augmented and virtual reality, and human-computer interaction.

HANDSIGHT: A TOUCH-BASED WEARABLE SYSTEM TO INCREASE INFORMATION ACCESSIBILITY FOR PEOPLE WITH VISUAL IMPAIRMENTS

by

Lee Stearns

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Jon E. Froehlich, Chair / Advisor
Professor Rama Chellappa, Co-Advisor
Professor Leah Findlater
Professor Ramani Duraiswami
Professor Gregg Vanderheiden, Dean's Representative

Dedication

To my family, who supported and encouraged me throughout this long process.

Acknowledgements

Above all else, I would like to thank the three professors with whom I have worked closely throughout this process. First, I thank my advisor, Jon Froehlich, for his guidance, support, and infectious enthusiasm. Thank you to my co-advisor, Rama Chellappa, for his patience and sense of humor. And thank you to Leah Findlater for her valuable advice and encouragement. I have learned a great deal from each of you about how to become a better researcher and would not have been able to complete this dissertation without you.

Thank you also to the two additional members of my dissertation committee, Gregg Vanderheiden and Ramani Duraiswami, for your time and valuable advice. Your input made this dissertation stronger, and your recommendations will help to improve the quality of my research and presentations in the future.

I use the pronouns “we” and “our” throughout this dissertation to acknowledge the contributions of other students, professors, and associates throughout nearly every stage of this research. I especially want to thank Uran Oh for working with me throughout much of the project. Thank you also to Ruofei Du, Liang He, Jonggi Hong, Alisha Pradhan, Anis Abboud, Victor De Souza, Alex Medeiros, Meena Sengottuvelu, Chuan Chen, Jessica Yin, Harry Vancao, Eric Lancaster, Catherine Jou, Victor Chen, Mandy Wang, Ji Bae, David Ross, Darren Smith, and Cha-Min Tang.

And thank you to my fellow lab mates for your advice, encouragement, and feedback: Matthew Mauriello, Kotaro Hara, Seokbin Kang, Dhruv Jain, Manaswi Saha, Majeed Kazemitabaar, Ladan Najafizadeh, and Brenna McNally.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AR	Augmented Reality
BLV	Blind and Low Vision
CMOS	Complementary Metal-Oxide Semiconductor
CNN	Convolutional Neural Network
CPU	Central Processing Unit
FAST	Features from Accelerated Segment Test
FOV	Field of View
FPS	Frames per Second
GPU	Graphics Processing Unit
HCI	Human-Computer Interaction
HMD	Head-Mounted Display
IMU	Inertial Motion Unit
IR	Infrared
KLT	Kanade-Lucas-Tomasi
LED	Light-Emitting Diode
LV	Low Vision
OCR	Optical Character Recognition
RGB	Red, Green, and Blue
RGBD	Red, Green, Blue, and Depth
SIFT	Scale Invariant Feature Transform
SVM	Support Vector Machine
VI	Visually Impaired
VR	Virtual Reality

Chapter 1: Introduction

Many activities of daily living such as getting dressed, preparing food, wayfinding, or shopping rely heavily upon visual information, and the inability to access that information can negatively impact the quality of life for people with visual impairments [16,72,108]. While previous research has explored solutions for assisting with visual tasks that can be performed at a distance, such as identifying landmarks for navigation [23,29,74,75,121,128,171] or recognizing people and objects [11,23,29,145,146,204], few have attempted to provide access to visual information through touch. Touch is a highly attuned means of acquiring textural and spatial information, especially for people with visual impairments [55,149]. By supporting touch-based access to information, we may help users to better understand how a surface appears (*e.g.*, document layout, clothing patterns), thereby improving the quality of life.

An assistive device that can detect touch events on physical surfaces and identify the content that is beneath the user's finger enables several potential applications, which can be subdivided into two categories: (1) access to visual information in the physical world, such as printed text, colors and textures, images, maps, and charts, and (2) control of computers or mobile devices to access digital information or specify application-specific commands. For the former, previous work has primarily focused on reading text on printed documents, product labels, or appliance displays [61,91,189,190]. Work in the second category has been more varied (*e.g.*, speech control [8], midair gestures [5,26]), but several approaches use touch-

based input performed on the body or other physical surfaces [25,63,64,70,184,228]. For example, *OmniTouch* [70] projects virtual controls onto the user's hand, arm, or a handheld object, and detects touch input using a shoulder-worn depth camera.

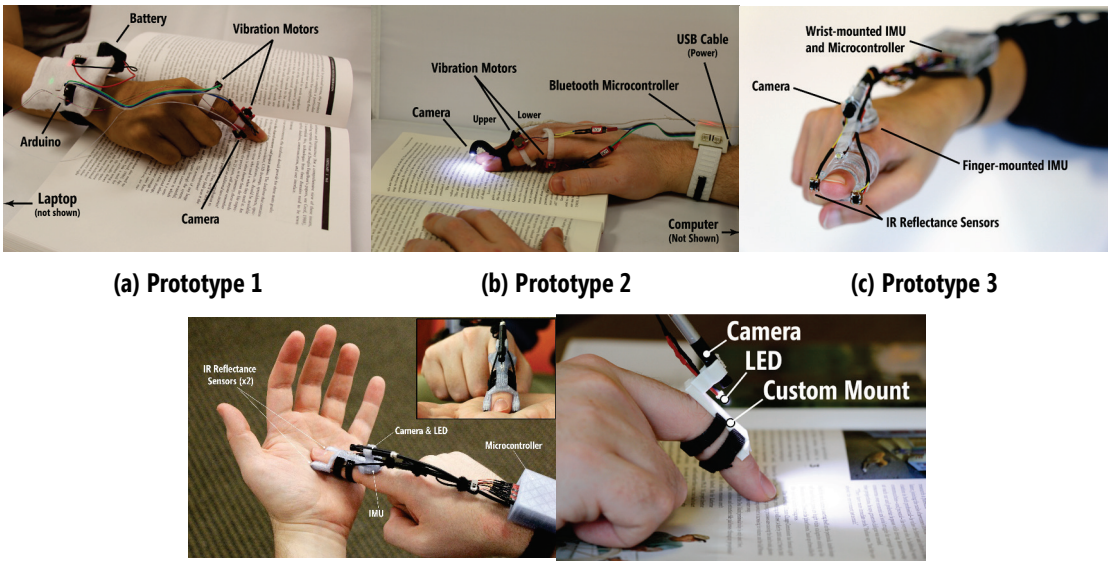
While researchers have begun to explore some aspects of touch-based information accessibility, several important open questions remain. Most prominent among these are the issues of sensing and feedback: *what is the best method to recognize the content the user is touching, and how should information about that content be conveyed to the user?* In particular, the location of the sensors plays a large role in the design of the physical system, algorithms, and user interactions—we hypothesize that finger-worn sensors will enable intuitive interactions and simplified algorithms. Furthermore, while touch-based interactions have several potential advantages, they also introduce new challenges—for example, accurately tracing a line of text while reading may be difficult for blind users. These issues must be addressed for a system that supports general touch-based access to information to be feasible.

1.1 Research Approach and Overview

To explore the potential benefits of accessing visual information through touch, the research in this dissertation focuses on augmenting a visually impaired user's finger with interactive, real-time computer vision to help them access information about the physical world. In particular, we present the design and evaluation of different prototypes for a system called *HandSight*, which uses wearable cameras and other sensors to detect touch events and identify surface content beneath the user's finger

(*e.g.,* text, colors and textures, images). There are four key aspects of HandSight: (i) designing and implementing the physical hardware, (ii) developing signal processing and computer vision algorithms, (iii) designing real-time auditory, haptic, or visual feedback that enables users with vision impairments to interpret surface content, and (iv) evaluating prototypes with visually impaired users to assess usability.

This dissertation describes several distinct but interrelated threads of research, each of which ties back to the core goal of supporting access to information through touch. We implemented and tested five proof-of-concept HandSight prototypes (Figure 1.1) consisting of a finger-mounted camera and other sensors. While our overarching goal is to increase the accessibility of information across a wide variety of settings, this dissertation focuses on three specific application areas: reading and exploring printed



(a) Prototype 1 **(b) Prototype 2** **(c) Prototype 3**

(d) Prototype 4 **(e) Prototype 5**

Figure 1.1: Five iterations of the HandSight prototype wearable finger-camera system. (a) and (b) show the iterations that were used for reading printed text, while (c) and (d) show the iterations that were used for detecting on-body input to control mobile devices and accessing digital information, and (e) shows a final iteration used for augmented reality magnification and for identifying clothing colors and visual patterns.

documents [197–200], controlling mobile devices through on-body interaction [202,203], and identifying colors and visual textures [135,201]. User studies demonstrate the feasibility of reliably recognizing several types of touch content, highlight strengths and weaknesses of our approach, and help uncover tradeoffs that will be important to consider when designing future wearable assistive devices.

1.1.1 Reading and Exploring Printed Documents

We first applied HandSight to reading and exploring printed documents. We conducted three studies to assess the feasibility of touch-based exploration and sequential reading. An important component of this feasibility evaluation was to determine if participants would be able to accurately follow a line of printed text in the absence of visual or tactile cues. In the first two studies [198,199], we used an iPad test platform to collect accurate finger-traces and isolate the interface from implementation details. We compared haptic and audio directional guidance and identified tradeoffs between the two conditions. Audio may result in slightly better line-tracing accuracy and be more familiar to users but could also distract from the synthesized speech content; haptic uses a different sensory channel and potentially offers clearer directional guidance but is less precise and may cause desensitization over time.

The third study [198] used a proof-of-concept finger-worn camera system to read physical documents and compared it with a state-of-the art smartphone application for reading printed text. Participants appreciated that our prototype provided immediate access to text content without the need to first capture the document, but overall they

preferred the fast and smooth text-to-speech output of the smartphone app. Ultimately, a finger-based reading approach may be best suited to material that is inherently spatial, such as maps or graphs, whereas existing applications that capture a global image of the document for text-to-speech may be preferred for text-heavy material.

In follow-up work [197,200], we extended our approach to assist low vision users with reading printed materials using augmented reality (AR) magnification. We conducted a series of design sessions with low vision participants to collect feedback on initial prototypes and solicit open-ended ideas about future wearable magnification aids. Our designs explored several virtual display options (*e.g.*, affixed to real objects *vs.* moving with the pointing finger), image acquisition approaches (head-mounted, finger-mounted, or smartphone), and interaction techniques (*e.g.*, voice commands, midair gestures, or touchscreen controls). Overall, participants liked the concept of AR magnification, especially the natural reading experience and ability to multitask afforded by the projected 3D displays. At the same time, our system also presented difficulties compared to participants' existing magnification aids, most notably a steeper learning curve and limitations of the AR hardware we used.

1.1.2 Controlling Mobile Devices with On-Body Input

Next, we applied HandSight to accessing digital information by controlling mobile devices. On-body input, which employs the user's own body as an interactive surface, offers several advantages compared to existing touchscreen devices. On-body taps and swipes provide lightweight and always-available control (*e.g.*, [63,70]), an expanded

input space compared to small-screen wearable devices like smartwatches (*e.g.*, [109,118,150,152]), and proprioceptive and tactile cues that can enable accurate input even without visual feedback compared to a touchscreen's smooth surface [64,154].

We investigated the feasibility of using finger-worn sensors to recognize on-body input and, in particular, the potential of location-specific, semantically meaningful contextual gestures (*e.g.*, tapping on the wrist to check the time or swiping on the thigh to control a fitness app). We conducted three studies to test this idea, two as offline algorithmic evaluations with sighted participants [202,203] and one in realtime with visually impaired participants [203]. For the first two studies we collected images and gestural data and performed offline experiments to test whether we could distinguish location-specific gestures on the body. We developed localization and gesture classification algorithms and evaluated their accuracy across the locations and gestures we had gathered. The high classification accuracies—above 95% on average for coarse body locations and gestures—demonstrated the feasibility of our approach.

In the third study, we implemented a realtime system with three distinct interaction techniques for performing common tasks (*e.g.*, checking the time, answering a phone call, or activating voice input). We then investigated the usability and potential of the realtime system with visually impaired participants. Our findings validated realtime performance with our target population and highlighted tradeoffs in accuracy and user preference across different on-body inputs. Participants' comments highlight positive reactions to on-body input as well as tradeoffs between the three interaction techniques. These tradeoffs reflect both algorithmic performance and

broader design implications. Our findings also highlight obstacles to robust on-body input recognition, especially for visually impaired users who cannot rely on visual cues.

1.1.3 Identifying Colors and Visual Patterns

Lastly, we applied HandSight to identifying clothing colors and visual patterns [135,201]. To assess feasibility, we collected two image datasets with various fabric patterns—one a small dataset collected using a finger-mounted camera representative of the items in a single user's closet, and one a much larger and more varied dataset assembled from online sources. We repurposed and fine-tuned state-of-the-art object classifiers to the task of fabric pattern classification, achieving high accuracy (99% and 92%) when training and testing with the first and second datasets individually. When training with the second, larger dataset and testing on the first—a much more difficult task, but one which demonstrates robustness and scalability—we achieve 73% accuracy, with most errors attributable to the finger-mounted camera's proximity to the fabric. We built an interactive prototype that positions the camera farther back on the user's finger to address this problem, and that also identifies the dominant fabric colors (*e.g.*, "striped blue and white").

This work is preliminary and primarily algorithmic, but it demonstrates feasibility and highlights the flexibility of a finger-based wearable device. Positioning the camera on the user's finger helps mitigate issues with inconsistent lighting and distance that can impact the accuracy of existing color and texture recognizers and allows for touch-based interrogation to better understand clothing appearance. Our

approach should allow users to quickly explore a surface and combine their sense of touch with visual texture and color information to make informed decisions about what to wear or buy.

1.2 Summary of Contributions

In summary, the overarching contributions of this dissertation are:

- Development and iterative refinement of HandSight, a novel wearable system to assist visually impaired users in their daily lives.
- Evaluation of HandSight across a diverse set of tasks, providing both empirical evidence and qualitative user feedback that demonstrate the advantages and disadvantages of a finger-worn approach in terms of physical design, algorithmic complexity and robustness, and usability.
- Identification of implications for the design of future wearable assistive systems and for the broader fields of accessibility, computer vision, augmented and virtual reality, and human-computer interaction.

This dissertation also makes specific contributions in four application areas, including:

- Implementation and systematic evaluation of haptic and auditory cues to assist blind users in following a line of printed text, identifying tradeoffs in terms of accuracy and user preference (Chapters 3 & 4).
- Exploration of the design space for augmented reality magnification and image enhancement, including proof-of-concept implementations evaluated and

refined through iterative co-design with low vision users and recommendations for future AR vision enhancement aids (Chapter 5).

- Offline algorithmic evaluations to test the feasibility of supporting on-body input using a finger-mounted camera and other sensors (Chapters 6 & 7).
- Design, implementation, and evaluation of a realtime on-body input system using finger- and wrist-worn sensors, with design reflections for on-body gestural interfaces in terms of what locations and gestures can be recognized most reliably across users (Chapter 7).
- Two novel fabric texture datasets, one collected systematically using a finger-mounted camera and the other assembled from fabric images downloaded from Google Images and augmented synthetically using rotations, scaling and cropping (Chapter 8).

1.3 Dissertation Outline

This dissertation is organized around three distinct applications of touch-based access to information. Chapter 2 provides background and related work. Chapters 3-5 explore the potential of using a finger-mounted camera to read printed materials; Chapter 3 describes preliminary work toward helping blind users to read through touch, and Chapter 4 builds on that work by investigating in greater depth questions related to reading using a finger-mounted camera and guiding a user's finger across a page. Chapter 5 considers the needs of low vision users, investigating augmented reality as a magnification aid. Chapters 6 and 7 apply the finger-mounted camera to help visually

impaired users control mobile devices and access digital information; Chapter 6 is a preliminary algorithmic investigation of the potential for localizing skin features from small image patches, and Chapter 7 explores the potential of using finger-worn sensors to recognize location-specific touch gestures on the user's skin and clothing. Chapter 8 applies the finger-worn camera system to identifying clothing colors and visual fabric patterns. And finally, Chapter 9 summarizes our findings and contributions and discusses opportunities for future work.

Chapter 2: Background and Related Work

This chapter covers background and related work for three areas of research most relevant to this dissertation. First, we survey academic literature and commercial products that use mobile or wearable cameras to assist users with visual impairments. We then focus on our specific goal of supporting touch-based information access, which is separated into two categories: access to visual surface information, and access to digital information (*e.g.*, by controlling a mobile device).

2.1 Portable Assistive Camera Systems

Rapid advances in camera technology and computer vision algorithms along with the ubiquity of mobile phone cameras have led to a wide variety of camera-based assistive devices for users with visual impairments [11,120,126,207,230,235]. For example, mobile and wearable cameras have been used as magnifiers for users with limited vision [35,170,235] and for both blind and low-vision users to support navigation and wayfinding [23,34,81], identification of faces [102,104], facial expressions [6,103], objects [23,29,81], and text on signs, products, or physical documents [163,187,189,190,230,233]. Applications that exploit mobile phone cameras are particularly appealing due to their affordability, portability, and adoption rate among visually impaired users [92,229]. However, mobile applications also require the use of one or both hands, limiting their availability while the user is otherwise occupied (*e.g.*, while walking with a cane or guide dog) [229]. Wearable systems benefit from being always available, potentially smaller, and more flexible in how they allow users to

interact with the device or external surfaces. These systems exist in a variety of form factors, including head-mounted (*e.g.*, glasses or headsets [104,128,132]), torso-mounted (*e.g.*, medallions, belts, backpacks [96,178,212]), wrist-mounted (*e.g.*, watches [194,195]), and finger-mounted (*e.g.*, rings [145,146,190]). Each of these designs offers advantages and disadvantages in terms of sensor flexibility and field of view as well as the user's sensitivity toward feedback mechanisms (*e.g.*, audio, haptic) that are co-located with the device. Mayol-Cuevas [133] and Velazquez [215] have written surveys that summarize these tradeoffs.

2.1.1 Smartphone Applications

Smartphone adoption among visually impaired individuals is nearly as high as it is for sighted individuals [229]. The ubiquity and accessibility of these devices means that assistive smartphone applications have the potential both to reach a large audience and to make a significant impact in the lives of visually impaired individuals [92]. Several applications apply the phones' camera hardware, processing power, and networking capabilities to help users read text, identify people or objects, or navigate indoor and outdoor environments. For example, *LookTel* [204] is an application that is designed to recognize money or user-customizable objects using the camera and scale-invariant (SIFT) features [122]. *KNFB Reader*[1] and *Text Detective*[2] are popular mobile applications that allow blind users to capture images of physical documents or signs,

[1] http://www.knfbreader.com/
[2] http://blindsight.com/textdetective/

which are parsed and read aloud using OCR and screen reader software. Apple[3] and Android[4] phones now include a built-in magnifier to assist low vision users and numerous free or low-cost third-party apps are available in the Apple or Google stores. *Seeing AI*[5] uses online servers to recognize and describe scenery, including text, currency, people, and colors. In contrast, *VizWiz* [11] does not perform automated recognition but instead sends images to paid crowd workers who can answer nearly any visual question (*e.g.*, reading text, identifying an object), providing greater flexibility and reliability—albeit at a slower rate, and with reduced interactivity and privacy [15]. *VizLens* [61] builds upon that work using computer-vision techniques for object and finger tracking to support interactive exploration of physical interfaces (*e.g.*, microwave buttons), while still benefiting from the reliability of crowd recognition and labeling. However, as a mobile phone application it still requires the use of one hand to hold the phone steady and aimed toward the target object, which could be challenging for blind users. Instead, our research uses wearable cameras to mitigate issues with aiming and to allow users to move both hands freely; we compare our work against smartphone applications in Chapters 4 and 5.

2.1.2 Cameras Worn on the Upper Body

While mobile phone applications are appealing because they use existing mass-market products and are therefore more affordable, reliable, and socially acceptable [92,192],

[3] iOS Accessibility: https://www.apple.com/accessibility/iphone/vision/
[4] Android Accessibility: https://support.google.com/accessibility/android/answer/6006949
[5] Seeing AI: https://www.microsoft.com/en-us/seeing-ai/

they are also limited to the types of sensing hardware and interactions available on the phone, require the use of one or both hands, and can be difficult to aim accurately toward the target object or content in the absence of sight [125]. Wearable cameras offer a potential alternative that could mitigate some or all of these issues. By mounting the camera on the head or chest, for example, a camera sensor will inherently face the same direction as the user and could therefore simplify the targeting process. Furthermore, a body-mounted camera leaves both of the user's hands free for performing other tasks, which is particularly important for use while navigating with a cane or guide dog.

A second iteration of the aforementioned VizLens [61] uses the Google Glass[6] head-mounted camera for capturing images and performing object and finger tracking. However, while the authors state that their prototype resulted in improved image quality through pilot testing, they did formally evaluate its usability. In contrast, a commercial product called *OrCam*[7] uses a glasses-mounted camera with speech feedback to recognize and read back text or to identify stored products and faces. In addition to processing complete images from the camera when users press a button, OrCam can also recognize pointing gestures to allow users to select a particular piece of content to be read aloud. User studies with legally blind and low vision participants are promising [140,217], but they have not evaluated the device's usability for totally blind users or tested the utility of spatial layout information for complex documents.

[6] https://www.google.com/glass/start/
[7] http://www.orcam.com/

Other researchers have explored the use of cameras mounted on the chest or shoulder to detect gestures performed midair or on nearby surfaces (including the user's own body). These types of gestures provide greater flexibility and availability than other mobile interfaces (*e.g.*, compared to touchscreens or voice input), and could allow visually impaired users to more easily interact with mobile devices to access digital information. For example, OmniTouch [70] used a shoulder-mounted depth camera to track the user's fingers and enable touch gestures on the palm, arm, or other surface alongside a small projector for visual feedback. They did not evaluate their device with visually impaired users and the interaction space was limited by the camera's field of view; however, their experiments and demonstrations are impressive and demonstrate the feasibility of supporting touch-based interactions on the body or other nearby surfaces using wearable cameras. Several other projects have also used cameras mounted on the chest to support recognition of midair [26,96,196] or on-body [63] gestures in order to control computers or mobile devices. In contrast, we use a finger-worn camera to provide similar touch-based content recognition capabilities and to support on-body interactions, while mitigating issues with framing the target content and supporting a wider and more flexible interaction space.

2.1.3 Head-Worn Vision Enhancement Systems

In Chapter 5, we explore a novel augmented-reality approach for vision enhancement using a head-worn system. Head-worn systems that include both a camera and display to enhance visual content are particularly promising for low vision users. Head-mounted displays (HMDs) for low vision users were first proposed in the 1990s

[66,130]. For example, the *Low Vision Enhancement System* (LVES, 1992; [130]) and the *Joint Optical Reflective Display* (JORDY, 1999; [45]) both used head-worn optics and displays to help low vision users magnify and enhance objects and text. Compared to other types of low vision aids, HMD-based solutions offer the potential advantages of portability, ready availability, and privacy while displaying enhanced information within the wearer's field of view. However, while some early work exists, HMD vision enhancement aids have only recently become truly feasible, and as such have been subject to very few empirical human-centered studies to assess their usability and potential. A recent study by Zolyomi *et al.* [237] showed that one such device (eSight [238]) improved access to information and social engagement but also had negative social impacts [237]. Another study by Profita *et al.* [168] investigated the social acceptability of HMDs, showing greater acceptance if the device is perceived as being used for an assistive purpose as opposed for a general mobile computing task.

Most HMD-based systems for low vision users project magnified and/or enhanced 2D video captured from a wearable camera onto screens mounted in front of the user's eyes [130,235,238–240]. Some recent examples have used consumer VR hardware: *ForeSee* [235] uses an Oculus Rift headset and *IrisVision* [240] uses a head-mounted smartphone (Samsung GearVR). Optical see-through displays have also been employed for vision enhancement [82,170], where virtual information is overlaid on a transparent display, thus augmenting rather than replacing the user's vision—the approach we take in Chapter 5. Google Glass has been used to display a magnified view of a smartphone screen [170] and to overlay enhanced edges onto the wearer's

view of the real world [82]; however, Glass itself is a low-resolution display (640×360) and not designed as a vision enhancement aid or augmented reality device (*e.g.,* the display is positioned in the user's visual periphery). In contrast, Zhao *et al.* [234] conducted an accessibility evaluation of the *Epson Moverio BT-200* smart glasses with participants with low vision. They concluded that while the semi-transparency of optical see-through displays did reduce contrast and make it somewhat harder for low vision users to read text or identify shapes, participants were able to successfully use the device and were positive about the experience, confirming that such devices are a useful prototyping platform for providing visual content to low vision users.

2.1.4 Cameras Worn on the Finger

As an alternative to cameras mounted on the upper body, the finger may offer several advantages. Especially for touch-based applications, moving the camera to the active finger may simplify the sensing algorithms, mitigate issues with framing or occlusion, and provide a higher-resolution view of the touched surface. Furthermore, for blind users the finger is a primary and highly attuned method for acquiring information during proximal tasks [55,149], and by augmenting the finger with additional sensing and feedback capabilities we may enable interesting and novel opportunities for touch-based interactions with the physical world. Finger-worn devices are becoming increasingly popular as sensors and processors continues to become smaller and more power-efficient, with the number of products, patents, academic publications, and conceptual designs increasing yearly (Table 2.1; see also the survey by Shilkrot *et al.*

System Name	Sensor type	Target Users	Application(s)	Proximity	Output
Merrill and Maes **(Unnamed)** [137]	IR transmitter and receiver	Sighted	Augmented reality interactions with physical objects.	Distant	Visual, speech
SmartFinger [172]	RGB camera	Sighted	Interact with digital devices, copy/paste data and images	Close/Touch	N/A
EyeRing [145,146]	RGB camera	Mainly BLV, also sighted	Identify currency, bar codes, copy and paste text	Close	Speech
Magic Finger [228]	RGB camera, optical mouse	Any	Recognize surface gestures, identify touched material	Touch	N/A
CyclopsRing [25]	RGB camera with fisheye lens	Any	Recognize midair & touch gestures, identify people/objects	Multiple	N/A
FingerSight [79]	Camera, laser	BLV	Recognize and convey edges using vibration cues	Distant	Haptic vibration
FingerReader [188–190]	RGB camera	BLV	Read printed text, music	Touch	Speech, Audio Cues, Haptic Vibrations
HandSight (our work)	RGB camera, IR, IMU	BLV	Read printed text, identify colors and patterns, recognize gestures	Touch	Speech, Audio Cues, Haptic Vibrations

Table 2.1: Overview of several recent finger-worn camera systems alongside our own work. "BLV" indicates that the system was designed for both blind and low vision users.

[191]). Ring form-factors are particularly appealing due to their potential for providing subtle, natural, and socially acceptable interactions and interfaces [177].

Several other researchers have explored applications of finger-mounted cameras, from reading to navigation to gestural input. One of the first instances of these was Merrill and Maes in 2007 [137], who used a finger-mounted infrared transmitter and receiver to enable distant augmented-reality interactions with physical objects. *SmartFinger* [172] and *EyeRing* [145,146] similarly enabled interactions with real-world objects through the use of a finger-mounted camera and simple computer vision algorithms, the latter also presenting a preliminary design of a shopping assistant application along with subjective reactions from visually impaired users.

Most relevant to our research are four recent projects: *Magic Finger*, *CyclopsRing*, *FingerSight*, and *FingerReader*. Magic Finger [228] used a small finger-mounted camera combined with an optical mouse sensor to enable touch gestures on any surface. In addition to detecting touch events and precise relative movements using

the optical mouse sensor, they were able to classify different types of surfaces using a tiny (1×1mm) RGB camera—potentially enabling context-specific interactions. While their evaluation was limited and their work was not intended specifically for visually impaired users, their methods for sensing touch locations and touch-based gestures strongly influenced our own finger-based approach.

In contrast, CyclopsRing [25] consisted of a camera with a wide-angle lens mounted between the fingers, which the authors used to recognize whole hand gestures (*e.g.*, pointing, pinching) or on-palm touch input (*i.e.*, drawing or performing gestures on the palm), as well as identifying objects that the user is pointing toward. It was not intended for users with visual impairments and was not evaluated formally; however, the example applications that the authors propose are promising and could extend to visually impaired users as well as sighted.

FingerSight [79] used a finger-mounted camera to enable blind users to interrogate the visual features of the surrounding environment through haptic sensory substitution. While they focused on distal interrogation, their methods could easily be applied to touch-based exploration of lines and other visual primitives. The authors concluded that their approach would enable blind users to differentiate between lines with an angular resolution of less than 15°, although they did not evaluate it with visually impaired users. This angular resolution somewhat contradicts our own findings in a related study with visually impaired participants [78], where we encountered a lower limit of ~23° for directional guidance. However, as our participants and methods were quite different, perhaps the two are not directly comparable.

Finally, FingerReader [188–190] used a finger-mounted camera alongside haptic and audio feedback to enable blind users to read printed text documents. They evaluated their prototype in two small studies with 3–4 blind participants, demonstrating the feasibility and technical capabilities of their approach. However, they did not report on any quantitative performance metrics, their participant preferences were conflicting, and all of the participants in their most recent study found it difficult to use FingerReader for reading printed text. These factors suggest that further investigation into the feasibility and usability of touch-based reading is necessary, especially in comparison to existing methods such as mobile scanners. We explore a similar touch-based reading approach in Chapters 3 and 4, evaluating our system with a larger number of participants—27 across three user studies—and focusing especially on methods for guiding the user's finger across the page while reading. Later work by Chu *et al.* extended FingerReader to read Chinese characters [31], while Shilkrot *et al.* built on their own work with FingerReader to read and play back sheet music [188].

2.2 Access to Visual Surface Information

By far the most important type of surface content for blind users to access is text. The inability to read menus, receipts and handouts, bills and other mail can negatively impact the daily activities of those living with visual impairments [16,72]. However, access to other visual surface information—such as colors and textures—are necessary for activities of daily living such as getting dressed or preparing food. Additionally,

access to lines, symbols, and images could help with understanding complex documents such as charts, tables, or maps. This section summarizes existing work toward improving the accessibility of this information in the physical world.

2.2.1 Reading Text using Optical Character Recognition (OCR)

Scientists have long sought to support blind people in reading printed text by developing new technologies (for reviews: [24,33,126]). Many early so-called "reading machines for the blind" used a sensory substitution approach where the visual signals of words were converted into non-verbal auditory or tactile cues. These systems were complicated to learn but increased the accessibility of printed text. Two such examples include the *Optophone* developed in 1914, which used musical chords or 'motifs' [37] and the *Optacon* (OPtical to TActile CONverter) from 1973, which used a vibro-tactile signal [12,54]. The Optacon continues to be used by blind readers, despite its slow reading speed and high learning curve, suggesting that these challenges are not necessarily a barrier to use.

With advances in sensing, computation, and OCR, modern approaches attempt to scan, recognize, and read aloud text in real-time. This transition to OCR and speech synthesis occurred first with specialized devices (*e.g., SARA CE*[8]*,* the original KNFB Reader[9]*,* [51]), then mobile phones (*e.g.*, *Text Detective*, KNFB Reader iOS), and now wearables (*e.g.,* FingerReader [189,190] and OrCam, described in the previous

[8] SARA CE: http://www.freedomscientific.com/Products/LowVision/SARA
[9] KNFB Reader Classic: http://www.knfbreader.com/products-classic.php

sections). Many of these devices and applications function as mobile scanners designed for capturing and processing documents under ideal conditions (*i.e.*, high contrast documents, simple fonts, good lighting and minimal perspective); however, others have begun to support recognition of text on signs and in natural scenes [47,186]. While decades of OCR work exist (*e.g.,* [28,142,187,219]), even state-of-the-art reading systems become unusable in poor lighting and require careful camera framing [86,127]. These limitations are true even for crowd-powered assistive applications such as VizWiz [11], VizLens [61], and *BeMyEyes*[10], which also introduce delays and negative implications for privacy compared to automated methods. Few existing systems provide access to spatial information that may be important for understanding content such as newspapers or menus. Two exceptions are OrCam, which supports basic pointing gestures to browse lines of text, and Kane *et al.*'s *Access Lens* [91], which is described in the next section. Compared to existing reading devices, our approach: (1) provides more intuitive and precise control over scanning and text-to-speech; (2) enables increased spatial understanding of the text layout; and (3) mitigates camera framing, focus, and lighting issues.

2.2.2 Identifying Colors and Patterns

Although text is the most common type of surface information that blind and visually impaired users need to access, other types of information can aid in various activities of daily living as well. Color and visual patterns are important for locating specific

[10] http://www.bemyeyes.org/

articles of clothing while getting dressed or shopping, and for coordinating outfits in a way that is visually appealing and socially acceptable [19,221]. This information is also important for distinguishing ingredients when preparing food (*e.g.*, green vs. red pepper), products on a shelf when shopping, or shaded regions when interpreting a graph or map.

Numerous commercial devices or smartphone applications have been designed to assist visually impaired users in identifying colors. For example, *Color Teller*[11] is a handheld device that users touch to an object, press a button, and then hear the recognized color aloud; *Color Star*[12] is also handheld and conveys color information through speech, but it functions at a distance and can also detect the presence of light sources, which it conveys through auditory cues or haptic vibrations. Smartphone applications such as *Color Identifier*, *Colored Eye*, and *Color Grab*—a few of the many that are currently available in the iOS and Android stores—function similarly but use the smartphone's camera and are much more affordable. While these products are highly beneficial and popular among visually impaired users, they have several important limitations. All are susceptible to the effects of ambient lighting and the distance from the target surface, especially the smartphone applications which must use the camera and flash that were not designed for up-close usage. Also, none support the recognition of textures or patterns, or efficient interrogation of multiple locations to assist in identifying multi-colored clothing and other objects.

---

[11] http://brytech.com/colorteller/
[12] http://www.caretec.at/Start.29.0.html

Showing promise for more advanced clothing pattern identification, Yuan *et al.* [209,227,231] developed a system to identify 4 patterns and 11 colors in images captured with a mobile phone or head-mounted camera. Blind users responded positively to the system although more detailed identification of colors and support for additional types of clothing patterns were desired. The interaction was also inefficient, requiring the user to hold out the clothing in front of them and use speech input to individually capture each still image to be classified. In contrast, Kane *et al.*'s Access Lens [91], which was primarily designed to enable touch-based access to physical documents, also included a color interrogation mode that allowed users to interrogate a document or object's color at arbitrary locations. However, Access Lens did not support recognition of textures or other visual primitives, and the authors did not present any findings from their evaluation of the color identification mode. Yang *et al.*'s Magic Finger [228] did support identification of textures for surface classification but did not focus on visually impaired users or consider ways to convey that information. Other researchers have explored haptic vibrations as a means to convey edges [79] or to identify colors and textures [18]; however, this research was preliminary and did not evaluate usability in practice, especially when combined with access to other information such as text. In Chapter 8, we apply our finger-mounted camera approach to identify clothing colors and visual patterns, allowing users to move their finger across an article of clothing and combine tactile information with continuous audio description of the fabric's appearance.

2.3 Access to Digital Information

While the above sections described work related to supporting access to visual information in the *physical world*, this section discusses the related task of accessing digital information from computers and mobile devices. The accessibility these devices has seen significant improvements in recent years, primarily due to advancements in touchscreen gestures, voice input, and screen reader technology. However, an HCI task that is simple for a sighted user may be much slower and more challenging for a blind user. For example, manually finding and playing a song can take 15 seconds for a blind user [89] while entering a four-digit passcode to unlock a smartphone requires on average eight seconds, leading many blind users to forgo this security feature altogether [9]. This disparity between sighted and blind or visually impaired users on common HCI tasks suggests that there is room for improvement. In this section, we summarize several techniques that have been proposed to make smartphones and wearable devices more accessible for users with visual impairments. We then survey alternate interaction techniques that use gestures on tables and other surrounding surfaces, or on the user's own skin and clothing. We apply these techniques as an input mechanism for HandSight to select between modes and to access digital information.

2.3.1 Smartphone and Smartwatch Accessibility

As mentioned in Section 2.1.1, smartphone adoption rates among blind and visually impaired users is very high [229], thanks in large part to advancements in speech recognition technology and the improved accessibility of touchscreens. The

combination of voice input and synthesized speech feedback (*e.g.*, Apple's *Siri*[13] or Google's *Assistant*[14]) provides a natural interface in the absence of visual information for dictating text, specifying commands, or requesting information. Indeed, previous research has shown that blind users tend to use this these features more frequently and for longer periods than sighted users [8]. However, speech input is not always possible due to concerns over privacy or social acceptability, and so it is important to support more discrete forms of input using the touchscreen or peripheral devices as well.

Although touchscreens have existed for decades, until relatively recently they presented a significant barrier to accessibility for people with vision impairments due to their reliance on visual cues and lack of tactile feedback. The ubiquity of smartphones and tablets that use a touchscreen as their primary input mechanism has brought a renewed interest in making touch interfaces accessible to all users [13,50,60,90]; some of this research has been incorporated into commercial products. For example, Kane *et al.*'s *Slide Rule* [90] interface allows the user to browse the screen's contents through multi-touch gestures and speech feedback in a manner that is very similar to Apple's *VoiceOver*[15] and Google's *TalkBack*[16] interfaces. While these interfaces significantly improve the accessibility of touchscreen devices, several limitations still exist [113,134]. The glass touchscreen does not offer much in the way of tactile feedback, which may limit the speed and accuracy of touch input for visually

[13] http://www.apple.com/ios/siri/
[14] https://assistant.google.com/
[15] http://www.apple.com/accessibility/iphone/vision/
[16] https://developer.android.com/design/patterns/accessibility.html

impaired users compared to alternative approaches [154]. Furthermore, for smartwatches and other wearable devices using small touchscreens for input, the size of the interaction space is severely limited and requires precise touch input that is challenging even for sighted users.

2.3.2 Touch Gestures on Arbitrary Surfaces

Numerous researchers have investigated the potential for augmenting physical surfaces to enable touch-based interactions with computers or home automation systems. For example, Rekimoto's *SmartSkin* interface [173,174] adds multi-touch capacitive sensing capabilities to tabletops and other surfaces using a grid of copper wires. Several other researchers [83,222,223,226] have explored the use of depth sensing cameras (*e.g.*, *Microsoft Kinect*) that are positioned above a table, and that can model the 3D geometry of a scene to recognize touch or midair gestures. The idea of augmenting a variety of physical surfaces to enable tangible touch-based interactions and create a larger interaction space is appealing, but as a general input mechanism it does not scale well; augmenting every surface with which a user might potentially wish to interact is simply not feasible. Sato *et al.*'s *Touché* system [184] requires only a single electrode to be attached to support capacitive touch sensing on nearly any physical surface, including tables, doorknobs, skin, and even water; however, even it cannot support interaction with arbitrary surfaces without prior modifications. Instead, if the user's body is augmented with self-contained sensing and feedback mechanisms, then the idea becomes much more tractable.

While a camera positioned above the input surface has a simpler calibration and sensing process, body-worn cameras can also support touch detection and gestural input in midair or on arbitrary surfaces. For example, Harrison et al.'s *OmniTouch* [70] uses a shoulder-worn depth camera and pico-projector to provide a portable touch display on a variety of surfaces including tabletops, walls, handheld notebooks, and the palm of the user's hand. Mistry and Maes's *SixthSense* [139] similarly uses a small RGB camera and projector worn around the neck as a pendant to enable interactions with arbitrary surfaces, although the absence of depth information prevents it from explicitly distinguishing touch gestures from midair pointing gestures.

As mentioned in earlier sections, finger-worn sensors have several potential advantages compared to those worn elsewhere on the body, including greater flexibility of input location and reduced problems with camera framing or occlusion. Despite these potential advantages, few researchers have applied them to sensing touch input on arbitrary surfaces. Kienzle and Hinckley's *LightRing* [97] uses a ring containing an infrared range sensor and gyroscope. Together, these sensors can detect touch events and recognize basic gestures, although their simplicity and positioning mean that they are not robust to unexpected finger movements and can recognize only relative motion. In contrast, Yang et al.'s Magic Finger [228] (discussed in Sections 2.1.4 and 2.2.2) combines a finger-worn optical mouse sensor for detecting touch events and track finger movement with a slower but higher-resolution camera to capture details for surface classification. These two sensors provide much more information enabling

more precise gesture tracking and localization, albeit at greater size and expense. Magic Finger also covers the user's fingertip, greatly reducing tactile sensitivity.

2.3.3 On-Body Input

On-body input provides several potential advantages over handheld or wearable touchscreen input (*e.g.*, smartphones or smartwatches) offering a larger input surface and more precise touch input even without visual cues [64,154]. However, how to sense this input and what form it should take are still open questions—which our research explores and partially addresses. Researchers have investigated a wide variety of wearable sensing approaches, including cameras [25,40,65,70,195,206,218], infrared [109,150–152], ultrasonic rangefinders [117,119], bio-acoustics [69,110], magnetic fields [27], electromyography (EMG) [131], electromagnetic phase shift [234], and capacitance sensors [117,131,184,220]. These approaches support a similarly wide variety of inputs, including discrete touches at different body locations [110,131,184], continuous touch localization on the hand or wrist similar to touchscreen input

System Name	Sensor type	Sensor placement	On body Interaction Space	Interaction type
OmniTouch [70]	Camera (Depth)	On the shoulder	On or above the hands or arms (limited by camera FoV)	Continuous touch locations
Touché [184]	Capacitive	Flexible (one on wrist, one elsewhere on body)	Flexible (requires the target location to be instrumented)	Discrete touch locations, body or hand pose
CyclopsRing [25]	Camera (RGB, Fisheye Lens)	Between fingers of passive hand (for on-body input)	On or above the instrumented hand	Continuous touch locations, touch gestures, hand pose
Botential [131]	EMG, capacitive	On the wrist (or arm, leg)	Flexible, different body parts	Discrete touch locations
ViBand [110]	Bio-acoustic	On the wrist	On the instrumented hand or arm	Discrete touch locations, non-directional gestures
SkinTrack [234]	Electromagnetic phase shift	On the wrist, ring on opposite hand	On the skin surface around the instrumented wrist	Continuous touch locations, touch gestures
WatchSense [195]	Camera (Depth)	On the wrist, facing toward fingers	On or above the instrumented hand (limited by camera FoV)	Continuous touch locations, touch and midair gestures
HandSight (Our Work)	Camera (Grayscale), IMU, IR	On top/side of the gesturing finger and wrist	Flexible (does not require additional instrumentation)	Discrete touch locations, touch gestures

Table 2.2: Overview of several recent on-body input approaches alongside our own work.

[25,70,195,234], and input based on 3D finger or arm positions [25,184,195]. We summarize a subset of this prior work alongside our own in Table 2.2, which helps to highlight the diversity of sensing approaches and on-body interactions. While these past approaches are promising, their sensor types and placements limit the types of interactions that they can support.

First, the interaction space is often constrained to a small surface (*e.g.*, wrist or arm) or to a narrow window in front of the user. Approaches using cameras mounted on the upper body (*e.g.*, [40,65,70,206]) restrict interactions to a pre-defined region within the camera's field of view. OmniTouch [70], for example, can only detect gestures on the hands or arms in a relatively small space in front of the user. Similarly, approaches using sensors mounted on one wrist or hand to detect gestures performed by the other hand (*e.g.*, [25,69,109,110,117,151,195,218,220,234]) limit on-body interactions to a relatively small area around the sensors. Some approaches such as Touché [184] or *iSkin* [220] are more flexible but still require instrumentation at the target interaction location, which limits scalability. In contrast, our approach places sensors on the *gesturing* finger, supporting input at a variety of body locations within the user's reach without requiring additional instrumentation. Further, our design could be readily extended to interact with surfaces beyond the body.

Second, prior work attempts to either identify touched body locations or detect motion gestures but not both. For example, Touché [184] and *Botential* [131] can localize touch input at various locations on the body using EMG or capacitance sensors. However, these systems cannot recognize relative surface gestures such as directional

swipes. In contrast, systems such as *PalmGesture* [218], *SkinTrack* [151], or *WatchSense* [195] can estimate precise 2D touch coordinates, enabling complex gesture interactions like shapes. However, these methods require sensors affixed on or near the interaction surface to achieve such precision, and they therefore cannot easily be extended to recognize multiple locations. Our approach uses a small finger-worn camera to identify touched locations, augmented by inertial and IR sensors for robust gesture recognition; together, these sensors enable location-specific gestures.

2.4 Summary

In this chapter we surveyed the academic literature and commercial products most relevant to our goal of supporting touch-based access to information for visually impaired users using wearable cameras and other sensors. We covered three active research areas, summarizing: (i) the current state of the art on mobile and wearable camera systems designed to assist visually impaired users, (ii) work toward increasing the accessibility of visual surface information (e.g., text, colors, and patterns), and (iii) work toward supporting access to digital information using mobile devices, and in particular using on-body input. We build upon this existing body of work in subsequent chapters, exploring issues related to the physical design, algorithms, and usability as we design and test HandSight.

Chapter 3: Reading Printed Materials by Touch: Initial Exploration

Despite the increased availability of digital information and screen reader software, reading printed text materials remains an important but challenging task for people who are blind or visually impaired. The inability to read menus, receipts and handouts, bills and other mail can negatively impact the daily activities of those living with visual impairments (*e.g.*, [16,72]). Although braille has long provided a promising alternative, fewer than 10% of the approximately 2 million adults with severe visual impairment in the United States are braille literate [147,148], and many materials are not available in braille format.

Although many devices and mobile applications—such as SARA CE[17], KNFB Reader iOS[18], and OrCam[19]—attempt to provide access to printed materials through camera capture and optical character recognition (OCR), open questions remain. One challenge is how to help blind readers properly aim the camera so that a target object is completely visible and centered within the camera's field of view (*e.g.*, [36,86,213]). To accommodate this issue, the popular KNFB Reader iPhone application, for example, provides a spoken report to describe whether the document is fully visible and rotated

---

This chapter contains work published in the proceedings of the 2nd Workshop on Assistive Computer Vision and Robotics (ACVR'14) in Conjunction with the European Conference on Computer Vision (ECCV'14) [199].

---

[17] SARA CE: http://www.freedomscientific.com/Products/LowVision/SARA
[18] KNFB Reader: http://knfbreader.com/
[19] OrCam: http://www.orcam.com/

correctly. Another challenge is how to interpret and communicate documents with complex layouts such as newspapers or menus. Determining which blocks of text to read, in what order, and what layout details to convey are known issues even with digital content [14,111].

Compared to mobile applications, our finger-based approach may mitigate overhead camera framing issues, enable a blind reader to better understand the spatial layout of a document, and provide better control over pace and rereading. A finger-based approach, however, also introduces new challenges that have not been fully investigated. Because the field of view from a finger-mounted camera is limited, the reader must precisely trace along the current line of text so that the image does not get cut off or distorted. Physical navigation through the document is also needed to support reading, such as finding the start of a text passage and moving from one line to the next. Thus, a finger-based reading approach is contingent not only on accurate text capture and OCR, but also on effective finger guidance.

As an initial exploration of finger-based sensing and feedback, we focused on the challenges associated with helping a blind user read printed text. At this stage, our research questions were primarily exploratory, spanning both the human-computer interaction (HCI) and computer vision algorithms: (i) How can we effectively guide the user's finger via haptic and auditory feedback to appropriately scan the target text and provide notifications for certain events (*e.g.*, start/end of line or paragraph reached)? (ii) How accurately can optical character recognition (OCR) be achieved at

a speed that is responsive to the user's touch? (iii) How does the position, angle, and lighting of the finger-mounted camera affect OCR performance?

We pursued two parallel approaches; to answer the computer vision questions, we developed an early HandSight prototype along with efficient algorithms for perspective and rotation correction, text detection and tracking, and OCR. This chapter presents preliminary evaluations and demonstrates the feasibility of our envisioned system. To answer the HCI-related questions, we developed a custom touchscreen-based test apparatus that simulated the experience of using HandSight but provided additional experimental control and allowed us to more precisely track the user's finger in response to feedback conditions. Using this setup, we report on a preliminary evaluation with four visually impaired participants (three blind) across three finger guidance conditions.

3.1 System Design

HandSight is comprised of three core components: sensors, feedback mechanisms, and a computing unit for processing. Our initial prototype is shown in Figure 3.1. Before describing each component in more detail, we enumerate our six design goals.

3.1.1 Design Goals

We developed the following design goals for our system based on prior work and our own experiences developing assistive technology:

1) **Touch-based interaction.** Although future extensions to HandSight could examine distal interaction, we focus on digitally augmenting the sense of touch.

2) **Should not hinder normal tactile function.** Fingers are complex tactile sensors [87,112] that are particularly attuned for people with visual impairments [55,149]; HandSight should not impede normal tactile senses or hand function.

3) **Easy-to-learn/use.** Many sensory aids fail due to their complexity and extensive training requirements [33]; to ensure HandSight is approachable and easy to use, we employ an iterative, human-centered design approach.

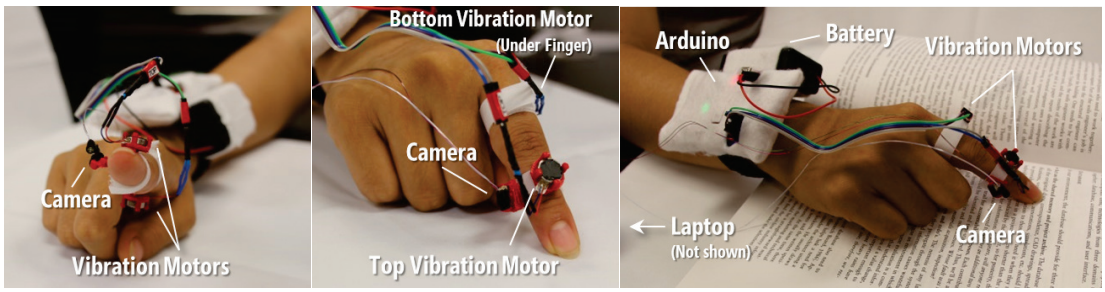
4) **Always available.** HandSight should allow for seamless transitions between its use and real-world tasks. There is limited prior work on so-called always-available input [143,182,183,228] for blind or low-vision users.

5) **Comfortable & robust.** HandSight's physical design should support, not encumber, everyday activities.

6) **Responsive & accurate.** HandSight should allow the user to explore the target objects (*e.g.*, utility bills, books) quickly—the computer vision and OCR algorithms should work accurately and in real-time.

3.1.2 Hardware

Our initial prototype used a small camera, vibration motors, and a laptop for processing and power. We describe each individual component below.



(a) Close-up front view **(b) Close-up side view** **(c) Full system view**

Figure 3.1: The initial HandSight prototype with a NanEye ring camera, two vibration motors, and an Arduino. Finger rings and mounts are constructed from custom 3D-printed designs and fabric. Processing is performed in real-time on a laptop (not shown).

Sensing Hardware. We use a single 1×1mm^2 *AWAIBA NanEye 2C* camera [7] that can capture 250×250 resolution images at 44 frames per second (fps). The NanEye was originally developed for minimally invasive surgical procedures such as endoscopies and laparoscopies and is thus robust, lightweight, and precise. The camera also has four LEDs coincident with the lens (2.5mm ring), which enables dynamic illumination control. The small size allows for a variety of finger-based form factors including small rings or acrylic nail attachments. In our current prototype, the camera is attached to an adjustable Velcro ring via a custom 3D-printed clip.

Processing. For processing, we use a wrist-mounted Arduino Pro Micro with an attached Bluetooth module that controls the haptic feedback cues. The video feed from the camera is processed in real time on a laptop computer (our experiments used a Lenovo ThinkPad X201 with an Intel Core i5 processor running a single computation thread at approximately 30fps).

Feedback. HandSight provides continuous finger-guidance feedback via vibration motors, pitch-controlled audio, or both. The initial prototype includes two

Figure 3.2: A demonstration of our perspective and rotation correction algorithm

vibration motors, 8mm diameter and 3.4mm thick (Figure 3.1). A text-to-speech system reads each word aloud as the user's finger passes over it, and distinctive audio and/or haptic cues can be used to signal other events, such as end of line, start of line, etc.

3.1.3 Image Processing Algorithms and Offline Evaluation

Our initial text detection and recognition algorithms involve a series of frame-level processing stages followed by between-frame tracking and merging once the complete word has been observed. Below, we describe our five stage OCR process and some preliminary experiments evaluating performance.

Stage 1: Preprocessing. We acquire grayscale video frames at ~40fps and 250x250px resolution from the NanEye camera (Figure 3.2). With each video frame, we apply four preprocessing algorithms: first, to correct radial and (slight) tangential distortion, we use standard camera calibration algorithms [71]. Second, to control lighting for the next frame, we optimize the LED intensity using average pixel brightness and contrast. Third, to reduce noise, perform binarization necessary for OCR, and adapt to uneven lighting from the LED, we filter the frame using an adaptive threshold in a Gaussian window; finally, to reduce false positives, we perform a

connected component analysis and remove components with areas too small or aspect ratios too narrow to be characters.

Stage 2: Perspective and Rotation Correction. The finger-based camera is seldom aligned perfectly with the printed text (*e.g.,* top-down, orthogonal to text). We have observed that even small amounts of perspective distortion and rotation can reduce the accuracy of text detection and OCR. To correct perspective and rotation effects, we apply an efficient approach detailed in [71,84,232], which relies on the parallel line structure of text for rectification. We briefly describe this approach below.

To identify potential text baselines, we apply a Canny filter that highlights character edges and a randomized Hough transform that fits lines to the remaining pixels. From this, we obtain a noisy set of candidate baselines. Unlikely candidates are discarded (*e.g.,* vertical lines, intersections that imply severe distortion). The remaining baselines are enumerated in pairs; each pair implies a potential rectification, which is tested against the other baselines. The pair that minimizes the baseline angle variance is selected and the resulting rectification is applied to the complete image.

More precisely, the intersection of each pair of baselines implies a horizontal vanishing point $V_x = l_1 \times l_2$ in homogeneous coordinates. If we assume the ideal vertical vanishing point $V_y = [0, 1, 0]^T$, then we can calculate the homography, *H,* that will make those lines parallel. Let $l_\infty = V_x \times V_y = [a, b, c]^T$ and calculate the perspective homography, H_p, using those values. The perspective homography makes

the lines parallel but does not align them with the x-axis. We must rotate the lines by an angle θ using a second matrix, H_r. The complete rectifying homography matrix is:

$$H = H_r H_p = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ a/c & b/c & 1 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ a/c & b/c & 1 \end{bmatrix} \quad (1)$$

To investigate the effect of lateral perspective angle on performance, we performed a synthetic experiment that varied the lateral angle from -45° to 45° across five randomly selected document image patches. The raw rectification performance is shown in Figure 3.3a and the effect of rectification on character-level OCR accuracy is shown in Figure 3.3b (the algorithm for OCR is described below).

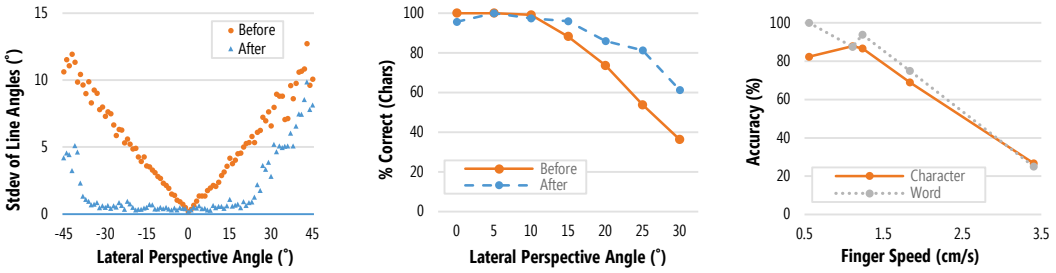
Stage 3: Text Detection. The goal of the text detection stage is to build a hierarchy of text lines, words, and characters. This task is simplified because we assume the perspective and rotation correction in Stage 2 has made the text parallel to the x-axis. First, we split the image into lines of text by searching for large gaps between text pixels in each row. Next, we split each line into words using an identical process on the columns of pixels. Gaps larger than 25% of the line height are classified as spaces between words. Finally, we segment each word into individual characters by searching for local minima in the number of text pixels within each column.

Stage 4: Character Classification. Real-time performance is important for responsive feedback, which prevents us from using established OCR engines such as Tesseract. Thus, we compute efficient character features (from [1]), and perform classification using a support vector machine (SVM). Each character candidate is centered and scaled to fit within a 32x32 pixel window, preserving the aspect ratio. The

window is split into four horizontal and vertical strips, which are summed along the short axis to generate eight vectors of length 32 each. These vectors, along with the aspect ratio, perimeter, area, and thinness ratio make up the complete feature vector. The thinness ratio is defined as $T=4\pi(A/P^2)$ where A is the area and P is the perimeter. We compensate for the classifier's relatively low accuracy by identifying the top k most likely matches. By aggregating the results over multiple frames, we boost performance.

Stage 5: Tracking and final OCR result output. The camera's limited field of view means that a complete word is seldom fully contained within a single frame. We must track the characters between frames and wait for the end of the word to become visible before we can confidently identify it. Character tracking uses sparse low-level features for efficiency. First, we extract FAST corners [179], and apply a KLT tracker [211] at their locations. We estimate the homography relating the matched corners using the random sample consensus [48]. After determining the motion between frames, we relate the lines, words, and individual characters by projecting their locations in the previous frame to the current frame. The bounding boxes with the greatest amount of overlap after projection determine the matches. When the end of a word is visible, we sort the aggregated character classifications and accept the most frequent classification. This process can be improved by incorporating a language dictionary model, albeit at the expense of efficiency. A text-to-speech engine reads aloud the identified word.

To investigate the effect of finger movement speed on OCR accuracy, we recorded five different speeds using a single line of text. The results are presented in



(a) Performance of lateral perspective and rotation rectification algorithm.

(b) Effect of lateral perspective angle on accuracy (before/after correction).

(c) Effect of finger speed on character- and word-level accuracy.

Figure 3.3: Results from preliminary evaluations of our (a-b) Stage 2 algorithms and (c) the effect of finger speed on overall character- and word-level accuracy.

Figure 3.3c. With greater speed, motion blur is introduced, and feature tracking becomes less accurate. In our experience, a "natural" finger speed movement for sighted readers is roughly 2–3cm/s. So, with the current prototype, one must move slower than natural for acceptable performance. Future iterations can compensate by using a higher frame rate camera (100fps) and by skipping frames as needed.

3.2 User Study to Assess Audio and Haptic Feedback

Our initial prototype implementation supported haptic and audio feedback, but how best to implement this feedback for efficient direct-touch reading is an open question. We planned to later conduct a user evaluation of the full system to assess the combined real-time OCR and finger guidance for a variety of reading tasks. At this initial stage, however, our goal was to refine the finger guidance component of the system through a preliminary evaluation of three types of feedback: (1) audio only, (2) haptic only, and (3) a combined audio and haptic approach. We conducted a user study with four visually impaired participants to collect subjective and performance data on these three

feedback types. To isolate the finger guidance from the image processing algorithms, we used a custom iPad app that simulated the experience of using the full system.

3.2.1 Method

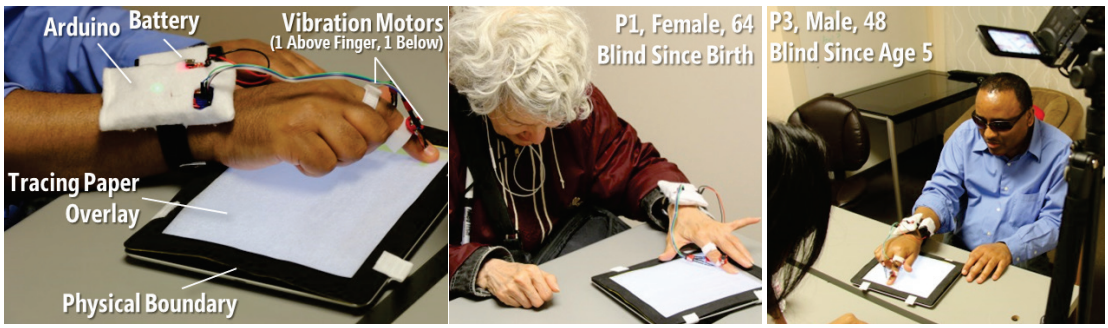
We summarize our experimental setup and methods below.

Participants. We recruited four VI participants; details are shown in Table 3.1. All participants had braille experience, and three reported regular use of screen readers.

Test apparatus. The setup simulated the experience of reading a printed sheet of paper with HandSight (Figure 3.4). It consisted of the hand-mounted haptic component of the HandSight system controlled by an Arduino Micro, which was in turn connected via Bluetooth to an Apple iPad running a custom experimental app. A thin foam rectangle acted as a physical boundary around the edge of the screen to simulate the edge of a sheet of paper, and the iPad was further covered by a piece of tracing paper to provide the feel of real paper and to reduce friction. The app displayed text documents, guiding the user to trace each line of the document from left to right and top to bottom. As the user traced their finger on the screen, text-to-speech audio was generated, along with the following feedback guidance cues: start and end of a line of text, end of a paragraph, and vertical guidance for when the finger strayed above or

ID	Age	Gender	Handedness	Level of Vision	Duration of Vision Loss	Diagnosed Medical Condition	Hearing Difficulties
P1	64	Female	Left	Totally blind	Since birth	Retinopathy of prematurity	N/A
P2	61	Female	Left	Totally blind	Since birth	Retinopathy of prematurity	Slight hearing loss
P3	48	Male	Right	Totally blind	Since age 5	N/A	N/A
P4	43	Female	Right	No vision one eye, 20/400 other eye	30 years	Glaucoma	N/A

Table 3.1: Background of the four user study participants.



(a) iPad test apparatus (b) Participant 1 (c) Participant 3

Figure 3.4: Study setup and test apparatus: (a) overview; (b-c) in use by two participants.

below the current line. Lines were 36 pixels in height and vertical guidance began when the finger was more than 8 pixels above or below the vertical center of the line.

Feedback conditions tested. We compared three finger guidance options, testing audio and haptic cues individually or in combination:

- *Audio only.* All guidance cues were provided through non-speech audio. The start and end of line cues each consisted of a pair of tonal percussive (xylophone) notes played in ascending or descending order, respectively. The end of paragraph sound was a soft vibraphone note. When the user's finger drifted below or above a line, a continuous audio tone would be played to indicate that proper corrective movement. A lower tone (300 Hz) played to indicate downward corrective movement (*i.e.,* the user was above the line). The pitch decreased at a rate of 0.83Hz/pixel to a minimum of 200Hz at 127 pixels above the line. A higher tone (500 Hz) was used to indicate upward corrective movement (up to a maximum of 600Hz with the same step value as before).

- *Haptic only.* The haptic feedback consisted of two finger-mounted haptic motors, one on top and one underneath the index finger (see Section 3.1.2).

Based on piloting within the research team, the motors were placed on separate phalanges so that the signal from each was easily distinguishable. To cue the start of a line, two short pulses played on both motors, with the second pulse more intense than the first; the reverse pattern indicated the end of a line. For the end of a paragraph, each motor vibrated one at a time, which repeated for a total of four pulses. For vertical guidance, when the finger strayed too high, the motor beneath the finger vibrated, with the vibration increasing in intensity from a low perceivable value to maximum intensity, reached at 127 pixels above the line; below the line, the top motor vibrated instead (again with the maximum intensity reached at 127 pixels).

- *Combined audio/haptic.* The combined condition included all of the audio *and* haptic cues described above, allowing the two types of feedback to complement each other in case one was more salient for certain cues than the other.

Procedure. The procedure lasted up to 90 minutes. For each feedback condition, we first demonstrated the feedback cues for the start/end of each line, end of paragraph, and vertical guidance. Next, we prepared a training article and guided the user through the first few lines. Participants then finished reading the training article at their own pace. Finally, we prepared a test article and asked participants to read through the text as quickly and accurately as possible. While we manually guided participants as necessary for the *training* article (*e.g.*, adjusting their finger), no manual guidance was provided during the *test* task. Four articles of approximately equivalent complexity were selected from *Voice of America* (a news organization), one for the training tasks

(a) Participant 1 finger trace (audio only condition) **(b)** Participant 1 finger trace (haptic only condition)

Figure 3.5: Our iPad test apparatus allowed us to precisely track and measure finger movement. Example trace graphs for Participant 1 (P1) across the audio- and haptic-only conditions are shown above (green is on-line; red indicates off-line and guidance provided). These traces were also used to calculate a range of performance measures. For example, for P1 the average overall time to read a line was 11.3s (*SD*=3.9s) in the audio condition and 18.9s (*SD*=8.3s) in the haptic condition. The average time to find the beginning of the next line (traces not shown above for simplicity but were recorded) was 2.2s (*SD*=0.88s) in the audio condition and 2.7s (*SD*=2.4s) in the haptic condition.

and one to test each feedback condition; all articles had three paragraphs and on average 11.0 lines (*SD*=1.0) and 107.0 words (*SD*=13.5). The order of presentation for the feedback conditions was randomized per participant, while the test articles were always shown in the same order. After each condition and at the end of the study, we asked questions on ease of use. We video recorded the sessions and logged all touch events.

3.2.2 Analysis and Findings

We analyzed subjective responses to the feedback conditions, and user performance based on logged touch events. Figure 3.5 shows a sample visualization from one participant (P1) completing the reading task in the *audio-only* and *haptic-only* conditions. Due to the small sample size, all findings in this section should be considered preliminary, but point to the potential impacts of HandSight and tradeoffs of different feedback types.

In terms of overall preference, three participants preferred *audio-only* feedback; see Table 3.2. Reasons included that they were more familiar with audio than haptic signals (P1, P3), and that it was easier to attend to text-to-speech plus audio than to text-to-speech plus haptic (P4). P2's most preferred condition was the *combined* feedback because she liked audio cues for line tracing and haptic cues for start/end of line notifications. In contrast, *haptic-only* feedback was least preferred by three participants. For example, concerned by the desensitization of her nerves, P1 expressed that: *"...if your hands are cold, a real cold air-conditioned room, it's [my tactile sensation] not going to pick it up as well."* P4 also commented on being attuned to sound even in the haptic condition: *"You don't know if it's the top or the bottom [vibrating] ...It was the same noise, the same sound."* As shown in Figure 3.6, ease of use ratings on specific components of the task mirrored overall preference rankings.

Participants were also asked to compare their experience with HandSight to braille, screen readers and printed-text reading using 5-point scales (*1-much worse* to

	Rank 1	Rank 2	Rank 3
P1	Audio	Combined	Haptic
P2	Combined	Audio	Haptic
P3	Audio	Haptic	Combined
P4	Audio	Combined	Haptic

Table 3.2: Overall preference rankings by participant. Audio feedback was the most positively received.

	Braille	Screen Reader	Printed Text
P1	3	3	3
P2	3	5	5
P3	4	4	4
P4	5	5	5

Table 3.3: Ratings comparing prior text reading experiences with HandSight; 1-much worse to 5-much better.

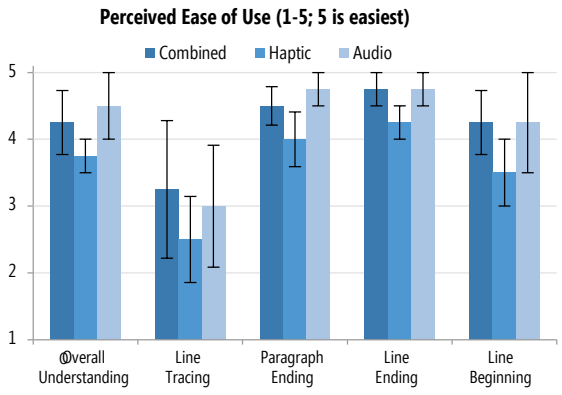


Figure 3.6: Average perceived ease of use of different text guidance attributes based on a 5-point scale (1-*very difficult*; 5-*very easy*). Error bars are standard error (*N*=4).

5-much better). As shown in Table 3.3, HandSight was perceived to be at least as good (3) or better compared to each of the other reading activities. In general, all participants appreciated HandSight because it allowed them to become more independent when reading non-braille printed documents. For example, P3 stated, *"It puts the blind reading on equal footing with rest of the society, because I am reading from the same reading material that others read, not just braille, which is limited to blind people only"*. P1, who had experience with Optacon [95], Sara CE, and other printed-text scanning devices also commented on HandSight's relative portability.

In terms of performance, we examined four primary measures averaged across all lines per participant (Figure 3.7): average absolute vertical distance from the line center, time spent off the line (*i.e.,* during which vertical feedback was provided), time from start to end of a line, and time from the end of a line to the start of the next line. While it is difficult to generalize based on performance data from only four participants, *audio-only* may offer a performance advantage over the other two conditions. *Audio-only* resulted in the lowest average vertical distance to the line center for all participants. Compared to the *haptic-only* condition, *audio-only* reduced the

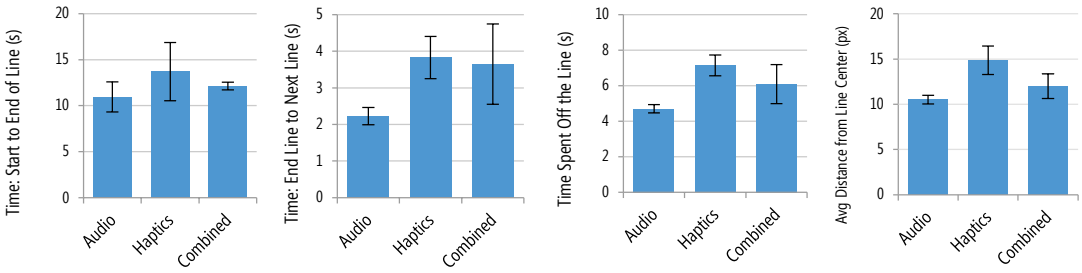


Figure 3.7: Average performance data from the four user study participants across the three feedback conditions. While preliminary, these results suggest that audio-only feedback may be more effective than the other options tested. Error bars show standard error; (*N*=4).

amount of time spent off the line by about half. It was also faster for all participants than *haptic-only* in moving from the end of a line to the start of the next line. We conduct a larger study in Chapter 4 to confirm these findings and to better assess what impact the feedback conditions have on reading speed from start to end of a line.

3.3 Discussion

Below, we discuss our preliminary findings and opportunities for future work.

Haptic Feedback. Though we have created many different types of finger-mounted haptic feedback in our lab, we tested only one in the user study: when the user moved above or below the current line, he or she would feel a continuous vibration proportional in strength to the distance from the vertical line center. Future work should experiment more with form factors, haptic patterns (*e.g.,* intensity, frequency, rhythm, pressure), number of haptic devices on the finger, as well as the type of actuator itself (*e.g.*, Figure 3.8). While our current haptic implementation performed the worst of the feedback conditions, we expect that, ultimately, some form of haptics will be necessary for notifications and finger guidance.

Blind reading. Compared to current state-of-the-art reading approaches, our long-term goals are to: (1) provide more intuitive and precise control over scanning and text-to-speech; (2) increase spatial understanding of the text layout; and (3) mitigate camera framing, focus, and lighting issues. Moreover, because pointing and reading are

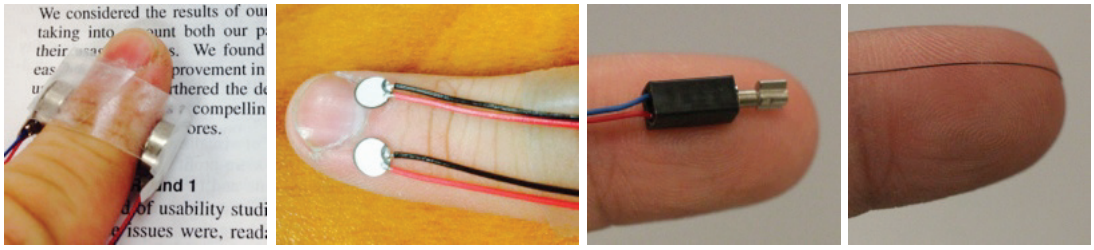


Figure 3.8: Haptic feedback alternatives: (a) 10×2.7mm^2 vibro-discs; (b) 5×0.4 mm^2 piezo discs; (c) 3×8 mm^2 vibro-motors; (d) 0.08mm Flexinol wire (shape memory alloy).

tightly coupled, finger-based interaction intrinsically supports advanced features such as rereading (for sighted readers, rereading occurs 10-15% of the time [94] and increases comprehension and retention [43,114]). We focused purely on reading text on simple documents, but in Chapter 4 we investigate more complex layouts so that the user can sweep their finger over a document and sense where pictures are located, headings, and so on. Future work should explore a variety of documents (*e.g.*, plain text, magazines, bills) and household objects (*e.g.*, cans of food, cleaning supplies), and examine questions such as: How should feedback be provided to indicate where text/images are located? How should advanced features such as re-reading, excerpting, and annotating be supported, perhaps through additional gestural input and voice notes?

Computer Vision. Our preliminary algorithms are efficient and reasonably accurate, but there is much room for improvement. By incorporating constraints on lower-level text features we may be able to rectify vertical perspective effects and affine skew. We can also apply deblurring and image stabilization algorithms to improve the maximum reading speed the system is able to support. Robust and efficient document mosaicking and incorporation of prior knowledge will likely be a key component for supporting a wider range of reading tasks.

Multi-sensory approach. Currently, our prototype relies on only local information gleaned from the on-finger camera. However, in the future, we would like to combine camera streams from both a body-mounted camera (*e.g.,* Orcam [159]) and a finger-mounted camera. We expect the former could provide more global, holistic information about a scene or text which could be used to guide the finger towards a target of interest or to explore the physical document's layout. We could also use the information to improve the performance of the OCR algorithms, by dynamically training the classifier on the page fonts and creating a generative model (*e.g.,* [123]).

3.4 Summary

Our overarching vision is to transform how people with VI access visual information through touch. Though we focused specifically on reading, this initial investigation offers a first step toward providing a general platform for touch-vision applications. The design and algorithmic evaluation of our initial HandSight prototype show the feasibility of our approach and highlight important technical issues that we must consider. Additionally, our user study, which evaluated three finger-guidance approaches using a controlled setup (the iPad test apparatus), found that, in contrast to prior work [189], haptic feedback was the *least* preferred guidance condition. The pitch-controlled audio feedback condition was not only subjectively rated the most preferred but also appeared to improve user performance. Clearly, however, more work is needed to explore this and other aspects of a touch-based approach to reading and exploring printed text materials, which we investigate in the next chapter.

Chapter 4: Evaluating Haptic and Auditory Directional Finger Guidance

Previous research—including the work described in the previous chapter as well as concurrent work by Shilkrot *et al.* [189,190] using a similar system called FingerReader—explored ring-based devices with embedded cameras that allow blind readers to trace their finger over printed text and hear real-time speech output. However, these studies focused on feasibility with small sample sizes (3–4 participants) and did not report on quantitative performance metrics. This prevents an in-depth understanding of finger guidance effectiveness, reading performance, and user reactions. The most recent of these studies underscores the need for further investigation: despite the theoretical advantages of finger-based reading, all three participants found it difficult to read text with FingerReader [190]. This provokes the question: *why*? To what extent are finger-based cameras a viable accessibility solution for reading printed text? What design choices can improve this viability?

To further investigate the feasibility of a finger-based sensing and feedback system for reading printed text, we conducted a controlled lab experiment to compare audio and haptic directional finger guidance with 19 blind participants using an iPad-based testbed (Study I). The primary goal was to compare the effects of the two guidance methods in terms of line tracing accuracy, reading speed, comprehension (through standardized comprehension questions), and subjective response. We later

---

This chapter contains work published in the ACM Transactions on Accessible Computing (TACCESS November 2016) [198].

(a) Camera Sensor (b) Finger-mounted Camera (c) Full Prototype

Figure 4.1: The first two iterations of the HandSight prototype use a 1×1mm^2 AWAIBA NanEye 2C camera developed for minimally invasive surgeries (*e.g.*, endoscopies) that can capture 250×250px images at 44fps (a). Also shown are two views of our finger-based reading system (b) and (c). Future designs can be made much smaller.

also randomly selected 4 of those participants to provide feedback on an updated HandSight wearable prototype (Figure 4.1), so as to help guide its design (Study II). These participants also provided feedback on the use of a smartphone app (KNFB Reader iOS) to read printed documents, which allowed us to compile a list of some of the relative advantages and disadvantages of each.

The findings from Study I showed similar performance between haptic and audio directional guidance, although audio may offer an accuracy advantage for line tracing. While a small majority of participants preferred haptic guidance to audio, the overall split reflects contradictions found in previous research [189,190,199]. Open-ended comments also highlight the tradeoffs of the two types of guidance, such as the interference of audio guidance with speech output and the potential for desensitization to haptic guidance. Finally, while several participants appreciated the direct access to layout information provided with HandSight's exploration mode, and the lower learning curve of HandSight as compared to braille, important concerns arose about ease of use and the amount of concentration required. In the follow-up sessions (Study II), while not offering a controlled comparison, participants appreciated that HandSight

provided immediate access to text content without the need to worry about first capturing the document, but overall they preferred the fast and smooth text-to-speech output of KNFB Reader iOS. Combined, these findings lead to new questions about finger-based reading, who may benefit the most from such an approach, and how to refine the design tested in our study.

The contributions of this chapter are: (1) empirical results comparing audio and haptic directional finger guidance for a reading task in terms of user performance and subjective response; (2) the implementation and preliminary evaluation of a real-time proof-of-concept system that combines a small finger-mounted camera and feedback mechanism with efficient computer vision algorithms to read printed text; and (3) design reflections for finger-based reading devices for people who are blind. While our long-term goal is to investigate the many interactions made possible by collocating sensing and feedback on the fingers, our focus here is on the interactions necessary to use such a system to explore and read a physical document.

4.1 Study I: Audio vs. Haptic Guidance for Finger-Based Reading

To investigate the exploration and reading of printed text documents using finger-based interactions, we conducted a controlled lab study with 19 blind participants. The primary goal of this study was to compare audio and haptic directional finger guidance methods in terms of user performance and preference. However, as the first larger-scale study of finger-based reading (N=19 *vs.* N=3 and N=4 [190,199]), the study also

quantitatively explored to what extent a finger-based reading approach can allow a blind reader to interpret the spatial layout of a document and to read and understand that document.

As in Chapter 3, we simulated the experience of reading a physical document using a touchscreen tablet (an iPad) covered with a sheet of paper (Figure 2c). This approach allowed us to bypass certain technical challenges in implementing a real-time camera and text recognition system, and instead to focus on the user experience of finger-based reading. The iPad also allowed us to collect precise finger traces to enable detailed finger-movement analysis not previously possible.

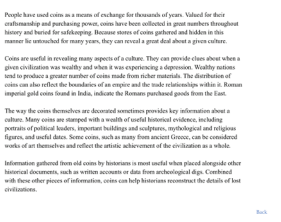
4.1.1 Method

In this controlled lab study, participants read two types of printed documents with audio and haptic finger guidance. We used a within-subjects design with a single factor of *Directional Guidance* that had two levels (*Audio* and *Haptic*); order of presentation of the conditions was fully counterbalanced. In addition to measuring reading speed and finger movement, we collected subjective feedback and assessed basic document comprehension using standardized questions. Despite similarities to Shilkrot *et al.*'s method [189,190], our protocol is an extension of our previous work [199], which was underway prior to the first FingerReader publication [189]. The final apparatus and method described here were also refined through pilot sessions with 5 additional participants (1 sighted, 1 low vision, 3 blind) who did not take part in the full study.

ID	Age	Sex	Vision Level	Braille		Screen Reader		Computer
				Use	Comfort	Use	Comfort	Comfort
P1	54	F	Blind	5	5	4	3	4
P2	33	F	Light	4	5	5	5	5
P3	55	M	Blind	3	5	5	5	4
P4	44	M	Light	2	2	5	5	5
P5	67	M	Blind	3	4	5	5	4
P6	62	M	Light	3	4	5	5	4
P7	40	M	Blind	1	1	5	4	4
P8	27	F	Light	5	5	5	5	4
P9	49	F	Light	5	5	5	5	3
P10	43	M	Blind	5	4	1	1	3
P11	44	M	Light	4	4	1	1	1
P12	39	M	Blind	4	5	5	5	5
P13	67	M	Blind	3	3	1	1	1
P14	50	F	Light	4	4	5	5	5
P15	26	M	Blind	5	5	5	5	5
P16	48	M	Blind	5	4	5	4	4
P17	59	F	Light	2	3	1	1	1
P18	47	F	Blind	4	3	1	1	1
P19	64	F	Light	4	3	4	4	3
Mean (*SD*)	48.3 (12.0)	N/A	N/A	3.7 (1.2)	3.8 (1.1)	3.8 (1.7)	3.7 (1.7)	3.5 (1.4)

Table 4.1: Study I participants. All participants were either blind or had minimal light perception (denoted "Light"). Frequency of use varied from 1 ("never") to 5 ("very often"), while comfort level varied from 1 ("very uncomfortable") to 5 ("very comfortable").

Participants. Twenty participants were originally recruited via campus email lists and local organizations, but one participant's data was discarded because he was unable to complete all of the required tasks. Of the remaining 19 participants, 11 were male and 8 were female, and the median age was 48 (*SD*=12.0, range 26–67). All participants were completely blind or had only minimal light perception. Five participants were congenitally blind, while the others had lost their vision later in life (some as recently as two years ago). As shown in Table 4.1, most participants were frequent users of braille, although 6 were just learning to read it and rated their comfort level as lower. All but 5 participants used screen readers at least some of the time and only 4 were not comfortable with computers and/or mobile devices. Participants were compensated for time and transportation.



People have used coins as a means of exchange for thousands of years. Valued for their craftsmanship and purchasing power, coins have been collected in great numbers throughout history and buried for safekeeping. Because stores of coins gathered and hidden in this manner lie untouched for many years, they can reveal a great deal about a given culture.

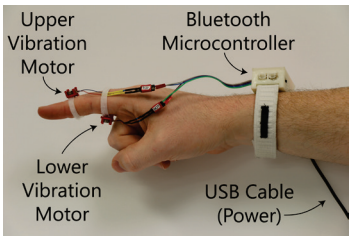
Coins are useful in revealing many aspects of a culture. They can provide clues about when a given civilization was wealthy and when it was experiencing a depression. Wealthy nations tend to produce a greater number of coins made from richer materials. The distribution of coins can also reflect the boundaries of an empire and the trade relationships within it. Roman imperial gold coins found in India, indicate the Romans purchased goods from the East.

The way the coins themselves are decorated sometimes provides key information about a culture. Many coins are stamped with a wealth of useful historical evidence, including portraits of political leaders, important buildings and sculptures, mythological and religious figures, and useful dates. Some coins, such as many from ancient Greece, can be considered works of art themselves and reflect the artistic achievement of the civilization as a whole.

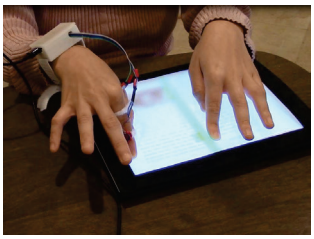
Information gathered from old coins by historians is most useful when placed alongside other historical documents, such as written accounts or data from archeological digs. Combined with these other pieces of information, coins can help historians reconstruct the details of lost civilizations.

Back

(a) Screenshot of iPad software showing a single-column document.



(b) Haptic feedback device, with actuator mounted on the finger.



(c) Test setup, with physical paper covering the iPad.

Figure 4.2: Study I test apparatus.

Apparatus. The test apparatus consists of an Apple iPad running custom software and connected via Bluetooth to a custom-built finger-worn haptic device (Figure 4.2). The source code is available on GitHub[20]. As noted previously, the iPad was used to provide a dynamic test environment that could precisely track finger movement in response to our directional guidance conditions. To simulate the feel of a physical document and reduce friction from the screen, a thin, blank paper covered the iPad. In addition, because there is no tactile border between the iPad screen and bezel, we added our own physical border made of 1/16" flexible foam (Figure 4.2c). The software displays documents and provides two modes of interaction: exploration and reading. All touch events (down, up, and move) on the screen are logged with x, y coordinates and timestamps.

Exploration mode. In this mode, audio cues allow users to gain a spatial sense of the document layout (*e.g.,* locations of images, columns, paragraphs) before transitioning to reading mode. As the user traces their finger over the document, they hear either a high-pitched flute sound when on a block of text or a low-pitched cello

[20] https://github.com/HCIL/HandSight

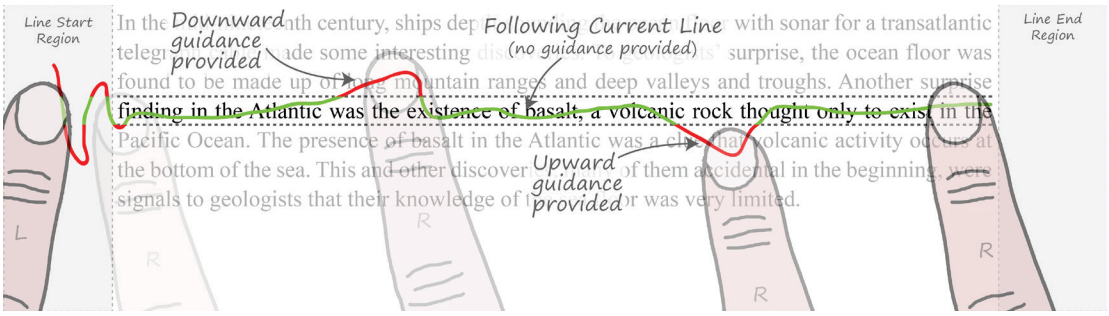


Figure 4.3: Reading mode interaction is bimanual. The user (1) places the right index finger in the "line start region" and moves vertically to find the start of the current line; (2) places the left index next to the right finger as an anchor; (3) traces the right finger along the line until it reaches the "line end region"; (4) returns the right index finger to beside the left finger before moving down to the next line. When the right finger is directly on the line (green trace) no directional guidance is provided, but when the finger moves too high or low (red trace), audio or haptic guidance indicates which direction to move to return to the line.

sound when on a picture. These sounds were selected and refined via pilot testing to be easily distinguishable by their pitch and timbre. When over whitespace, such as between paragraphs or columns, no sound plays.

Reading mode. In this mode, the user traces their finger from left to right along each line of text, while the system generates text-to-speech output using Apple's default iOS speech synthesis engine and provides directional finger guidance (haptic or audio depending on the condition). Reading is bimanual: the left hand, which is uninstrumented, serves as a line anchor (see "line start region" in Figure 4.3) while the right index finger traces the line. To begin reading, the user moves their right finger to the line start region shown in Figure 4.3 and an audio cue of ascending xylophone notes plays. If the finger is not already at the first line of text, audio or haptic feedback guides the user's finger up or down. Once the right finger is properly positioned over the "line start region" of the first line, the left hand joins the right hand and subsequently serves as a line anchor.

The user then traces his/her finger along the line to the right, while the system speaks each word aloud and provides audio or haptic guidance whenever the finger strays above or below the line (Figure 4.3). The speed of the text-to-speech output adapts to match the speed of the finger movement. Speech is provided only for the current line, and only when the user's finger is within 73 pixels (0.7cm) of the middle of the line (simulating a finger-mounted camera's field of view). At the end of the line ("line end region"; Figure 4.3), another audio cue plays, this time with descending xylophone notes, and the text-to-speech stops. The user then moves their finger left again to find the line start region and read the next line in the same manner. Finally, at the end of a paragraph, a new audio chime plays. The audio cues for the start and end of line and end of paragraph were selected to be easily distinguishable, which we again verified using early feedback from pilot participants.

For *audio directional guidance*, the system provides a continuous tone that varies in pitch. A low pitch indicates that the finger should move downward, and a high pitch indicates that the finger should move upward. If the finger is properly positioned over the current line, no audio plays. If the user's finger moves above the line, an audio tone at frequency 300Hz begins playing. If the user's upward movement continues, the frequency linearly decreases based on distance, down to a minimum of 200Hz at 127 pixels (1.2 cm). The 200Hz tone continues for any movement more than 2.4cm above the line. Similarly, if the user's finger strays below the line, the audio frequency begins at 500Hz and increases to a maximum of 600Hz at 127 or more pixels away. The choice

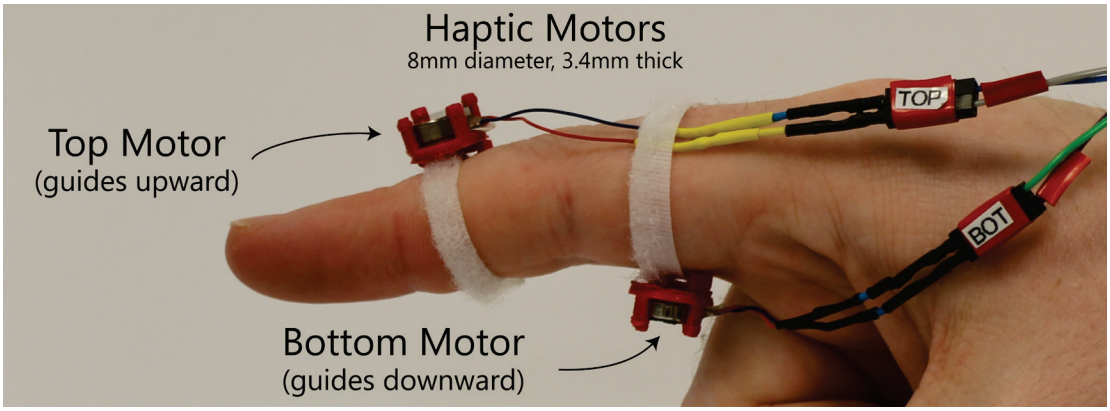


Figure 4.4: Close-up view of the haptic motors mounted on the finger via Velcro rings. The top motor vibrates when the user's finger moves below the line, providing upward guidance; the bottom motor vibrates when the user's finger moves above the line, providing downward guidance. The intensity of vibration depends upon the distance to the line, achieving maximum intensity at 127 pixels (~1.2 cm).

to vary audio frequency to indicate distance and direction was motivated by our prior work [155,199], and the exact pitches and thresholds were selected after pilot sessions.

The *haptic directional guidance* includes two vibration motors (8mm diameter disc, 3.4mm thick) controlled by an Arduino Pro Micro that communicates with the iPad via Bluetooth. The motors are attached to the user's right index finger with separate Velcro rings (Figure 4.4), one on top of the finger on the intermediate phalange and one below the finger on the proximal phalange. The lower motor indicates that the finger should move downward, and the upper motor indicates the opposite. Neither motor vibrates while the user's finger is directly over the current line of text. Vibration intensities off the line range from a minimum perceptible strength to the maximum strength the motors can provide, using the same distance thresholds as the audio condition. The choice to vary the position and intensity of vibration to indicate direction and distance was also motivated by our prior work [199] and validated in pilot sessions.

In early testing within our research lab and with external pilot participants, we tested multiple mappings for audio and haptic cues and intended finger direction (*e.g.*, higher pitch to indicate up *vs.* the opposite). Users were split in terms of which mappings were most intuitive, a point we revisit in the Discussion (Section 4.3.1).

Procedure. Each study session lasted 1.5–2 hours. Throughout, we employed two document types (Figure 4.5): single-column plain text, and two-column magazine-style with a figure and an article heading. For the reading tasks described below, we adapted four test documents from a Grade 8 Iowa Test of Basic Skills practice book [167]. The original text was modified slightly for length and to ensure clarity with our speech synthesis engine (*e.g.,* removing unnecessary proper nouns); see Appendix. The documents were thus all at similar reading levels and had multiple-choice comprehension questions. We also created training documents that were similar in length to the four test documents.

Following a background questionnaire, participants first learned how to use the document exploration mode as a precursor to the more complex task of both exploring and reading a document. The experimenter demonstrated the audio cues for text and images in exploration mode, then asked participants to explore one plain document and

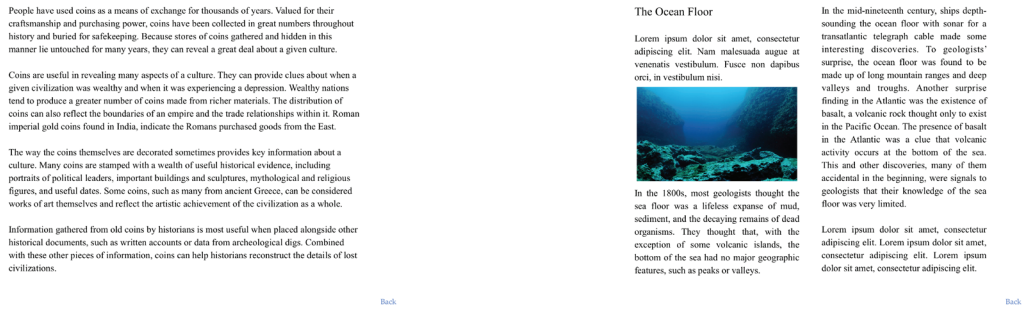


Figure 4.5: Examples of our test documents: plain text (left), and magazine (right).

one magazine document for up to three minutes each. To ensure that participants understood the exploration mode, we asked questions about the structure and layout of each document (*i.e.*, how many paragraphs and columns, are there pictures or headings and if so where are they located). To avoid biasing participants toward a particular exploration strategy or interpretation, we initially provided very little direction aside from demonstrating the audio cues and warning participants of the questions they would be expected to answer. After recording the answers for a document, the experimenter then guided participants to find the correct answers to ensure that they could later use the mode correctly prior to each reading task.

After the introduction of the exploration mode, participants explored and read documents with each of the directional guidance conditions (audio and haptic). The order of presentation for these conditions was fully counterbalanced. Document order was identical across all participants so that the documents themselves were matched an equal number of times with each guidance condition. To ensure similar physical experiences across conditions, participants wore the Arduino wristband and finger rings with the haptic motors throughout the full study session.

The procedure for each directional guidance condition was identical, with training using a plain document (~10 minutes) followed by testing with two documents (one plain and one magazine). For the training document, the experimenter demonstrated the feedback cues and participants incrementally learned to follow a line, find the next line or paragraph, and listen to the speech feedback while moving their finger. For each test document, participants were allowed up to 90 seconds in

exploration mode to assess the layout before the experimenter switched the system to reading mode. For the plain document, the reading task was to locate the first line of text and read the entire document. For the magazine document, participants read the last paragraph in the first column and the first paragraph in the second column. Exploration mode was used to locate the start of text for each document, as well as the start of the second column for the magazine document. After each test document, two multiple-choice comprehension questions provided in the Grade 8 Iowa Test of Basic Skills practice book were administered. At the end of each guidance condition, participants were asked about subjective ease of use. Finally, at the end of the study participants were asked to compare the two directional guidance conditions. See Appendix B for the full text of the subjective questionnaires.

Before conducting this study, we validated our selection of test documents and comprehension questions in a simple baseline study. Ten sighted college-age participants listened to synthesized speech of the four test documents and the comprehension questions. All 10 participants answered the questions correctly.

Data and Analysis. Collected data included log files from the iPad, participant responses to close- and open-form questions, and experimenter observations. To compare reading performance with haptic and audio guidance, we examined the following subtasks separately:

- *Line finding:* Finding the start of the current line. A line finding instance began with the first right-handed touch within the line start region (Figure 4.3) and ended with the finger exiting that region. Sometimes participants'

search paths resulted in more than one exit from the start region, so we included all data up to the final exit. For each instance, we calculated elapsed time and, as an error measure, the length of the movement path traced.

- *Line tracing:* Tracing left-to-right along the current line. A line tracing instance included all touch points after a successful line finding subtask until the right index finger entered the line end region (Figure 4.3). For each instance, we calculated reading speed in words per minute (wpm), and, as an error measure, the average absolute distance of the finger from the vertical center of the line across all x-coordinates in that line trace.

- *Full document:* Reading the full document from the start of the first line to the end of the final line. This comprehensive analysis includes all line finding and line tracing subtasks for a single document, as well as the time to transition between columns for the magazine documents. For each document, we calculated the average reading speed in words per minute (wpm) as well as the number of skipped words that were not read aloud.

Across the 19 participants, we collected data for 1513 lines. We identified outlier samples that were more than 3 standard deviations away from the mean for a given participant and condition, removing 31 samples (2.0%) of line tracing subtask samples and 49 (3.4%) of line finding subtask samples.

We used paired t-tests to compare line tracing speed between haptic and audio guidance. However, other measures violated the normality assumption of a t-test

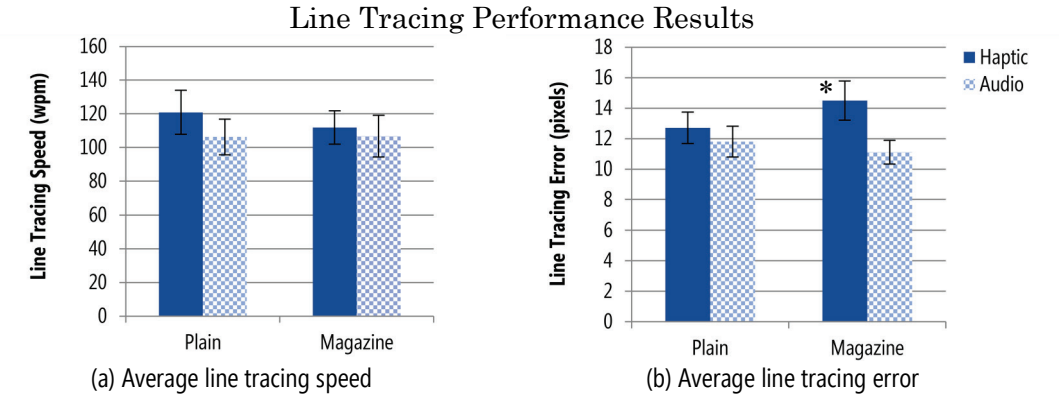
(determined using separate Shapiro-Wilk tests for each measure, $p < 0.05$). For these measures, we conducted non-parametric Wilcoxon signed rank tests to compare haptic and audio. For all posthoc pairwise comparisons, we applied Holm's sequential Bonferroni adjustments to protect against Type I error [77].

4.1.2 Findings

Our findings include quantitative performance results derived from the log data and exploratory qualitative descriptions of how participants responded to and interacted with the finger-based reading approach (*e.g.*, initial use of exploration mode, potential advantages of such an approach).

Reading Mode—Line Tracing. Figure 4.6 shows line tracing performance. For plain documents, the average reading speed with haptic guidance was 120.9 wpm (*SD*=57.0), compared to only 106.3 wpm (*SD*=46.2) with audio; however, a paired t-test comparing the two types of guidance was not statistically significant. A similar trend followed for magazine documents, at 111.8 wpm (*SD*=43.3) and 106.7 wpm (*SD*=54.1) for haptic and audio, respectively, with a paired t-test revealing no statistically significant difference between the two.

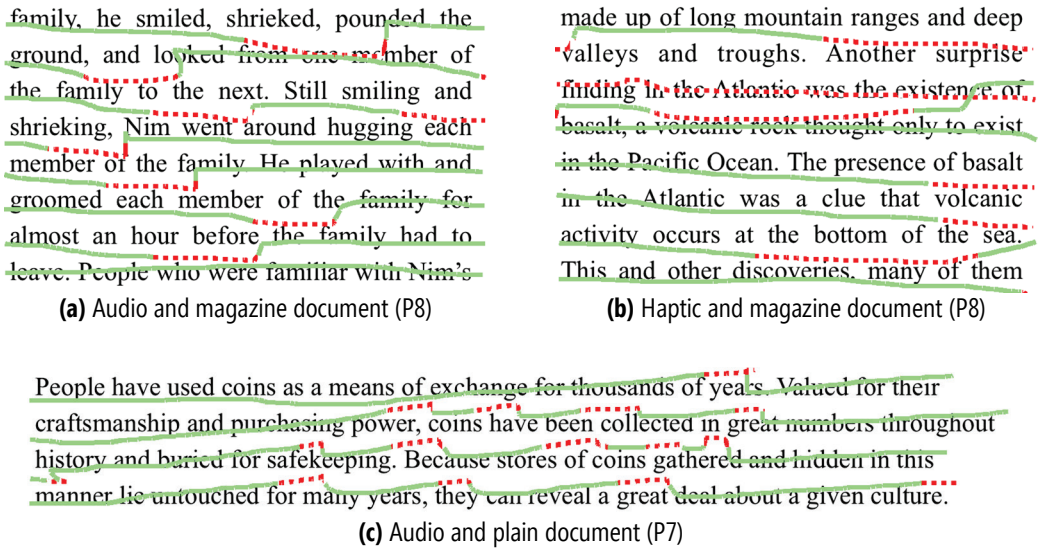
In terms of error, audio guidance was significantly more accurate than haptic guidance for the magazine documents, with an average distance of 11.2 px (*SD*=3.5) to the center of the line versus 14.6 px (*SD*=5.7). A Wilcoxon signed rank test was statistically significant on this measure, with a large effect size (Z_{19}=-2.374, p=.018, r=.54). Figure 4.7 shows a representative finger trace that illustrates this performance



(a) Average line tracing speed

(b) Average line tracing error

Figure 4.6: Average line tracing speed (higher is better), and average error—vertical distance offset from the center of the line (lower is better). Error bars indicate standard error (*N*=19). Performance was generally similar between the audio and haptic conditions, but audio resulted in significantly lower line tracing error for the magazine document (*).



(a) Audio and magazine document (P8)

(b) Haptic and magazine document (P8)

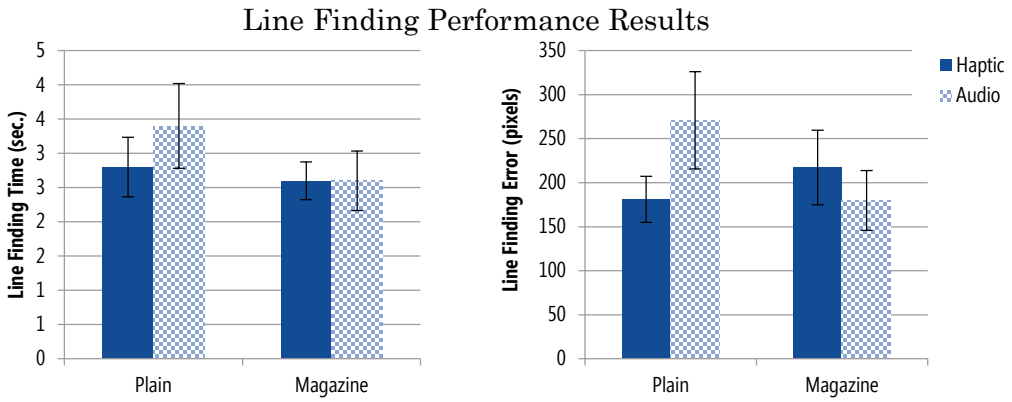
(c) Audio and plain document (P7)

Figure 4.7: Example finger traces. Solid (green) indicates that the finger was on the line, while dotted (red) indicates that the finger was off the line and directional guidance was being provided. (a) and (b) illustrate the difference in accuracy between the audio and haptic guidance conditions for P8. Participants frequently reacted more immediately to audio guidance but tended to ignore small amounts of vibration with haptic guidance. This observation may explain the significant difference in error between the audio and haptic conditions. Participants also tended to drift consistently above or below a line as they read, as seen in (a), (b) and (c).

difference. For the plain documents, however, the two guidance conditions resulted in more similar distances, at 11.9 pixels for audio (*SD*=4.6) and 12.8 pixels for haptic (*SD*=4.6). This difference was not significant using a Wilcoxon signed rank test.

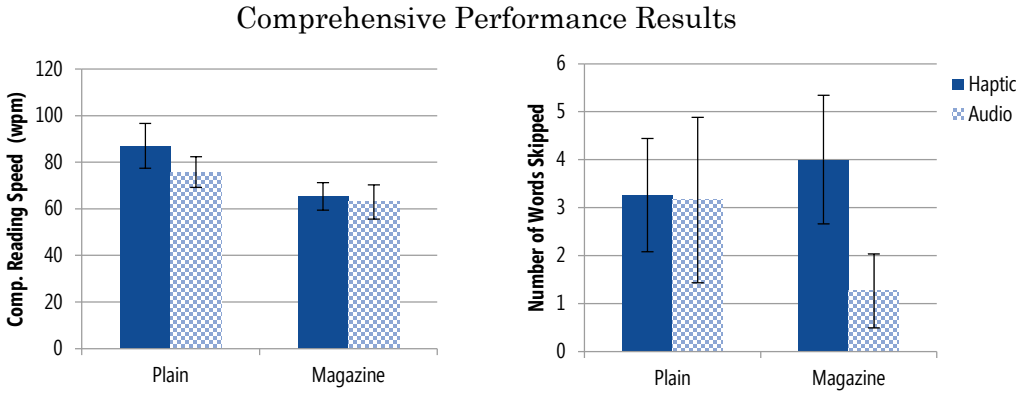
Participants tended to drift frequently, spending on average 29.7% (*SD*=13.2) of their line tracing time off of the line for the audio condition and 37.8% (*SD*=14.7) for the haptic condition. Reflecting the distance accuracy results above, this difference was statistically significant with a Wilcoxon signed rank test (Z_{19}=-2.57, p=.010, r=.59). In addition, participants tended to drift consistently above or below the line. Figure 4.7a, for example, illustrates downward drift whereas Figure 4.7c shows upward drift. We observed 11 participants who drifted consistently upward, 4 who drifted consistently down, and 4 who varied by document or did not tend toward either direction. This tendency may have been affected by how each participant's arm was positioned relative to the iPad—participants were instructed to rotate the screen as needed, but few chose to do so.

Reading Mode—Line Finding. As shown in Figure 4.8, line finding performance was similar across all directional guidance conditions and document types. No significant differences were found between haptic and audio guidance for either document type or performance measure using Wilcoxon signed-rank tests. Across all conditions, it took participants on average 2.6–3.4 seconds to find the next line in a document (plain: haptic *M*=2.8 seconds, *SD*=1.9, and audio *M*=3.4, *SD*=2.7; magazine: haptic *M*=2.6, *SD*=1.7, and audio *M*=2.6, *SD*=1.9). For error, measured as the average path length while searching for the start of a line, haptic averaged 181.2



(a) Average line finding speed **(b) Average line finding error**

Figure 4.8: The average time elapsed (left) and error (right) in finding the next line; lower is is better for both graphs. The error bars indicate standard error (*N*=19). Performance differenceses between the two conditions were not significant.



(a) Average comprehensive reading speed **(b) Average number of skipped words**

Figure 4.9: The comprehensive reading speed for an entire document (higher is better) and total number of skipped words (lower is better) by document. The error bars indicate standard error (*N*=19). Performance differences between the two conditions were not significant.

pixels (*SD*=114.0) with plain documents, while audio averaged 270.9 pixels (*SD*=241.1). In contrast, for magazine documents, haptic averaged 217.1 pixels (*SD*=184.4), compared to 179.8 pixels for audio (*SD*=147.5). Again, however, these differences were not found to be statistically significant.

Reading Mode—Overall Performance and Comprehension. Figure 4.9 shows comprehensive performance over the documents, including the total reading

time and number of skipped words. Reading speeds ranged from 63–81 wpm (plain: haptic *M*=81.1 wpm, *SD*=42.1 and audio *M*=75.8 wpm, *SD*=29.0; magazine: haptic *M*=65.4, *SD*=25.6 and audio *M*=63.0, *SD*=31.9). Overall, the number of skipped words, that is, words that were not read aloud by the text-to-speech engine, was uniformly low across conditions. The four documents contained an average of 211.5 words, but only 1–5 of those words were skipped on average for any given document. The number of skipped words was also similar between conditions for the plain documents (plain: haptic *M*=3.3, *SD*=5.2 and audio *M*=3.2, *SD*=7.5; magazine: haptic *M*=4.0, *SD*=5.9 and audio *M*=1.3, *SD*=3.3). Using Wilcoxon signed-ranks tests, no significant differences were found between haptic and audio guidance for either measure (speed, number of skipped words) with either document type.

While further investigation is needed to determine to what extent audio and haptic guidance impact comprehension, overall, participants answered the comprehension questions with high accuracy. Across all participants and conditions, 85% of the questions were answered correctly (Table 4.2).

Overall Subjective Response. Overall preference was split, with a small majority of participants (11 out of 19) preferring haptic feedback, 7 preferring audio, and 1 reporting equal preference. Participants also rated the two types of guidance in terms of comprehension and line tracing ease, from *1 – very difficult* to *5 – very easy*. The ratings, shown in Table 4.3a, support the overall preference patterns. Both guidance conditions were rated somewhat positively for both measures (3.1 or higher on average), and the differences between the two conditions were not statistically

Guidance	Document	2/2 Correct	1/2 Correct	0/2 Correct
Audio	Plain	14 participants	3 participants	2 participants
Haptic	Plain	17	2	0
Audio	Magazine	12	5	2
Haptic	Magazine	14	5	0

Table 4.2: Number of participants who answered the set of two comprehension questions correctly in each experimental condition (*N*=19). Most questions were answered correctly regardless of condition.

		Question	*N*	Mean	*SD*
(a)	Ease of use: Haptic vs. audio	Reading comprehension with audio guidance	19	3.2	1.3
		Reading comprehension with haptic guidance	19	3.7	1.2
		Line tracing with audio guidance	19	3.3	1.3
		Line tracing with haptic guidance	19	3.1	1.4
(b)	Ease of use: Elements common to both conditions	Start of text detection	19	4.5	0.7
		Start of line detection	19	4.2	0.6
		End of line detection	19	4.7	0.5
		End of paragraph detection	19	4.6	0.8
		Start of column detection	19	3.8	1.2
(c)	Comparison to existing technologies	HandSight vs. braille	18	3.0	1.0
		HandSight vs. screen readers	14	2.9	1.2
		HandSight vs. other reading aids	12	2.4	1.2

Table 4.3: Study I subjective ratings from 1 to 5 where 5 is best. (a) Reading comprehension and line tracing for each guidance condition. (b) Experience with subtasks common to both guidance conditions. (c) Overall comparison (better/worse) of HandSight versus braille, screen readers, and other reading aids. A score of 5 indicates that HandSight was perceived as much better than the existing technology, while a score of 1 indicates that it was much worse.

significant with Wilcoxon signed rank tests. Some challenges with the HandSight approach were seen as common to both types of guidance. For example, P12 said, *"The haptic feedback only tells you when you're not in line, not where the next thing would be"*, and made a similar comment for audio guidance.

The 11 participants who preferred haptic guidance generally felt that it was more intuitive, easier to use or faster than the audio. For example, P13 stated: *"It gave me a clearer indication of which way, up or down"*. P9 also commented, *"The vibrations kind of helped as a prompt, so that I automatically would go in the right direction, and I was able to read faster"*. Six of the participants who preferred haptic guidance also mentioned that the audio guidance was more distracting, and that made

it harder to focus on the speech feedback: *"You could focus on the audio of the text, and not be listening for other sounds"* (P7), or *"I missed a couple words because I was being distracted by the [audio]"* (P15). Even 4 of those who preferred audio guidance mentioned that the overlapping sounds could be somewhat distracting.

Of the 7 participants who preferred audio, almost all (N=6) found haptic guidance to be confusing: *"Sometimes when I use the vibrations I would forget which direction I was going based on where the vibration was"* (P5), or *"I had to analyze more what the vibrations meant"* (P14). Two participants also mentioned concerns about comfort, especially for prolonged use, for example: *"If you're reading longer your finger might get numb and it might get more difficult to figure out where the vibration was"* (P14).

Participants found the audio cues common to both guidance conditions relatively easy to use. Using these audio cues to detect the start of the text, line start/end areas, start of a column, and end of paragraph were all rated above 3.8 on a 5-point scale (Table 3b). Detecting the start of a column received the lowest score (M=3.8), perhaps reflecting the challenge of reading text with a more complex layout. This challenge can be non-trivial for some users. It should be noted that the participant whose data we discarded (described in Section 4.1.1, under "Participants"), had been blind since early childhood and was thus unfamiliar with the concept of a two-column document, an issue that requires further consideration in future work. He asked: *"Can a document be structured this way, with a paragraph just taking half part of the page?"*

Other participants also found the magazine document to be more difficult, especially those who were congenitally blind, but all were able to successfully complete the task.

Comparison with Other Technologies. As shown in Table 4.3c, the overall experience of HandSight was rated similarly compared to braille (*M*=3.0, *SD*=1.0), and somewhat negatively compared to other aids such as cell phone apps or scanner hardware (*M*=2.9, *SD*=1.2), and screen reader software (*M*=2.4, *SD* = 1.2).

Seven participants who were not comfortable with braille or existing reading technologies generally liked the lower learning curve and flexibility of our reading approach. For example, P11, who was currently learning braille said: *"With braille you gotta always constantly remember which dots are for which letters […]. this will tell you what the word is. Less stress."* (P11). P7 also commented on the utility of being able to directly control reading speed with our approach: *"A [screen] reader you get like one speed, it doesn't slow down for any reason, and sometimes it's a lot harder to go back and get your place from where you stopped."*

However, nine participants who were more familiar with braille and other reading devices raised concerns about ease of use and cognitive load. P14, for example, preferred braille: *"Reading braille I can read at a steadier pace and I can know where the punctuation is, and it's easier for me to find the next line"* (P14). Both P16 and P18 commented on cognitive load: *"There's the need to concentrate on staying within lines"* (P16), and, *"I'm so focused on trying to read the document, I'm not necessarily retaining the information the way I want to"* (P18).

Initial Use of Exploration Mode. The analyses above focus on reading mode, but at the start of the study, participants first used exploration mode to receive feedback on the presence of text, images and whitespace in both plain text and magazine documents. Even with this initial use, all but one participant correctly identified the presence or absence of a picture in both documents and described the picture's location. Determining whether audio breaks represented a gap between two paragraphs or two columns was more difficult, such that 11 participants initially identified multiple columns in the plain text document. However, between the two documents, the experimenter revisited how to distinguish between paragraphs and columns, and almost all participants (17 out of 19) were able to report the correct number of columns for the magazine document. Precisely counting paragraphs was still difficult, with only 9 and 3 participants reporting the correct number for the plain and magazine documents, respectively. For the magazine document the primary source of error was confusion over the definition of a paragraph in a multi-column document—the majority (N=15) did not count the paragraphs in the two columns separately. Additionally, 7 participants mistook the heading in the magazine document for another paragraph, and only 9 answered questions about it correctly.

We observed a few exploration strategies, with some participants using multiple strategies: 8 initially moved their fingers quickly but in no discernible pattern, searching out the locations of images and text within the document; 8 followed a procedure similar to reading braille, exploring left to right sequentially down the page; 12 explored sequentially left to right then top to bottom, counting breaks in the sound

to identify paragraphs and columns. Though we only told participants that we would ask them about the number of paragraphs, columns, and the presence/location of certain features (*e.g.,* headings, pictures), 6 participants provided additional details such as the width of the margins and the size and locations of the images and blocks of text.

Four participants provided unprompted feedback that they liked the document layout knowledge provided by the exploration mode. P6, for example, compared this advantage of the finger-based approach to a traditional screen reader:

> *"You have a perspective of the document layout—how many columns, where the graphics are located, the heading, and things like distribution of the text itself. […] When you use screen readers, you don't have any idea about that, you just get the text, you just get the content, but you don't have any direct access or idea of the document layout" (P6)*

P15 was particularly excited about the idea, using the exploration mode to identify the size and locations of images and blocks of text, and speculating based on their relative positions that *"maybe [this block of text is] a description of the picture. I always wonder things like that."* In contrast, P12 stated that he didn't see a use for spatial information in most situations: *"Not for blocks of text, but […] for diagrams or for maps it might be, because that's the only time that you actually need spatial orientation on a page."* He felt that a system that could automatically process a page and abstract the layout would be preferable. Further investigation is needed to evaluate how much this additional spatial information impacts comprehension or document

understanding, as well as how to best present that information to the user via audio or haptic feedback. We return to this point in the Discussion section.

Summary of Study I Findings. Audio and haptic guidance resulted in relatively similar user performance, although audio may offer an accuracy advantage for line tracing with some documents (it was significantly better than haptic for the magazine document). Although the majority of participants preferred haptic guidance, the overall split in preference reflects contradictions found in previous research [189,190,199]. Open-ended comments also highlight the tradeoffs of the two types of guidance, such as the interference of audio guidance with speech output and the potential for desensitization with haptic guidance. Finally, while several participants appreciated the direct access to layout information provided with HandSight's exploration mode, and the lower learning curve of HandSight compared to braille, important concerns arose about ease of use and the amount of concentration required.

4.2 Study II: Preliminary Use of a Proof-of-Concept Prototype

Following the in-depth comparison of audio and haptic finger guidance in Study I, we recruited 4 participants to return and provide qualitative feedback on a proof-of-concept wearable prototype. These follow-up sessions allowed us to collect preliminary evidence of: (1) the extent to which a blind reader can use a finger-mounted camera and directional guidance system to explore and read a printed document, and (2) the strengths and weaknesses of finger-reading versus a mobile scanner and screen reader.

4.2.1 Method

Participants explored and read printed documents using a proof-of-concept finger-mounted camera system, followed by KNFB Reader iOS, a popular mobile document reader. This was not intended to be a controlled comparison of the two technologies, but instead allowed for preliminary user experience feedback.

Participants. We randomly selected 4 participants from Study I to return for this follow-up study, with the constraint that they represent a mix of preferences for haptic and audio directional guidance. Study II was conducted shortly after Study I was completed, with participants returning between 1 and 3 weeks after their initial session. Participants' durations of blindness varied from 2 to 30 years, but none were congenitally blind. Only one participant (P12) had experience with KNFB Reader iOS. Refer to Table 4.1 for demographic information and to Table 4.4 for experience with specific technologies, including KNFB Reader iOS. As with Study I, participants were compensated for their time and transportation costs.

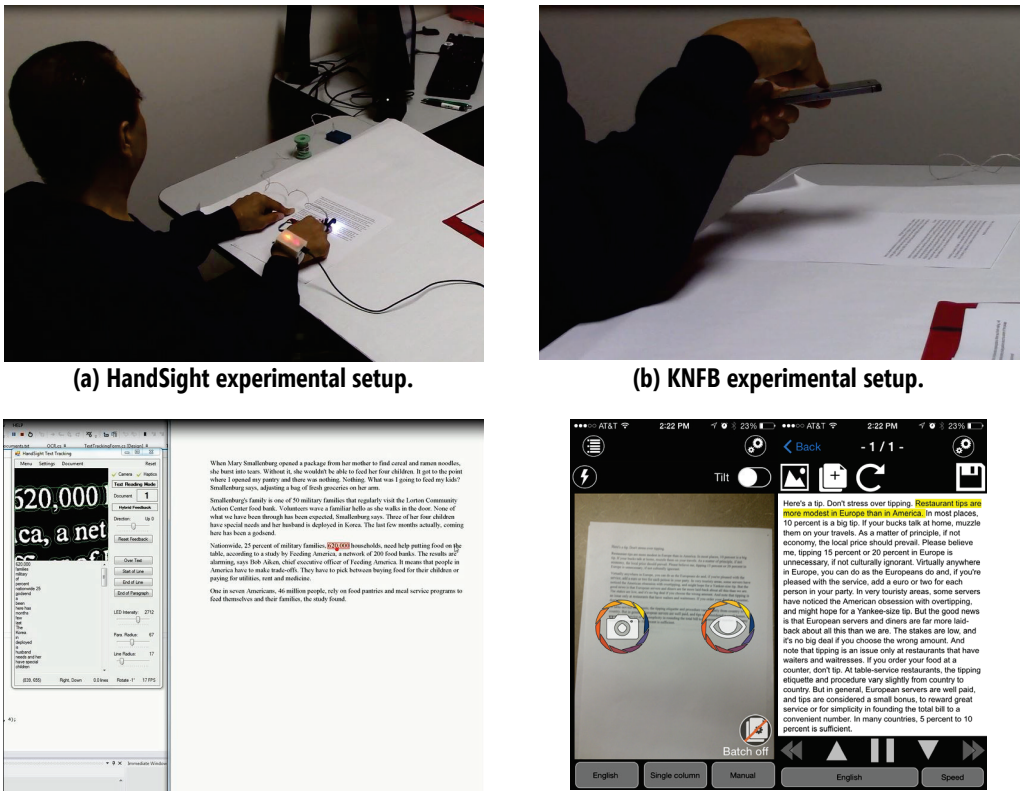
Apparatus. The proof-of-concept HandSight prototype consisted of a desktop computer running custom software, external speakers, a finger-mounted camera, and the haptic device from Study I (Figure 4.10). The camera was a self-illuminated Awaiba NanEye 2C CMOS camera and LED ring (~40 fps, 90° square field of view, 250x250

ID	Study 1 Feedback Preference	Frequency of Braille Use	Frequency of Screen Reader Use	Familiar with KNFB Reader iOS?
P10	No Preference (Tested Audio)	5	1	No
P11	Haptic	4	1	No
P12	Audio	4	5	Yes
P19	Haptic	4	4	No

Table 4.4: Study II participants; IDs are carried over from Study I. Comfort levels ranged from 1-5, with 1 indicating "very uncomfortable" and 5 indicating "very comfortable".

pixels, 2.4mm diameter), embedded in an adjustable ring and positioned above the finger to point down at the page (Figure 4.1b). The camera was positioned 1–2cm above the page and had a field of view approximately 1.5cm across (2–3 lines of text). These numbers varied somewhat depending on the participant's hand position.

As with Study I, the software provided two modes of interaction: *exploration* and *reading*. Exploration mode provided the same feedback as in Study I, except that



(a) HandSight experimental setup. **(b) KNFB experimental setup.**

(c) Screenshot of HandSight software. **(d) Screenshots of KNFB Reader iOS.**

Figure 4.10: Study II experimental setup. (a) The HandSight test apparatus consisted of a desktop computer running a custom reading program, stereo speakers, a finger-mounted camera system, and the haptic feedback device from our first study. Participants were asked to read through two documents using our prototype system. (b) The KNFB experimental setup consisted simply of an iPhone with the KNFB Reader iOS app. Participants were asked to read three documents using the app. (c) A screenshot of HandSight's OCR interface (this was not shown to the participant and used only by the experimenter). (d) Two screenshots of KNFB Reader iOS: (left) the 'capture' interface helps users orient the phone's camera to take a photo of the target document; (right) the digitized document screen-reading interface.

the prototype system did not detect images; as such, documents used in Study II did not include images**.** To ensure that the flute sound did not stop between individual characters or lines of text, the system first blurred the text using a blur radius that was manually calibrated prior to beginning the exploration tasks. The audio and haptic cues in reading mode were identical to those in Study I, with text-to-speech output using the IVONA Voice for Windows speech synthesis engine**.**[21] Exploration and reading events were logged with timestamps, but we could not log precise finger-trace data as we had done with the iPad in Study I.

The software processed each video frame from the camera using OpenCV[22], an open-source computer vision and image-processing library. With each frame, we applied four preprocessing algorithms. First, to correct radial distortion from the camera lens, we used standard camera calibration algorithms [71]. Second, to reduce noise, perform binarization necessary for OCR, and adapt to uneven lighting from the LED, we filtered each frame using an adaptive threshold in a sliding window. Third, to reduce false positives, we performed a connected component analysis and removed components with areas too small or aspect ratios too narrow to be characters. Finally, to correct for finger rotation, we blurred the image to efficiently group the components into likely lines of text, then extracted the minimum-area bounding rectangle for each new component. We used the estimated orientation of this rectangle to correct for

[21] http://www.ivona.com/us/for-individuals/voices-for-windows/
[22] http://opencv.org/

camera rotation, inverting it so that the lines of text were parallel to the *x*-axis. This process is similar to that described in Chapter 3 [181].

To simplify sensing for this proof-of-concept prototype, we assumed that a complete image of the page was available to the system in advance. The software then estimated the current finger location by performing OCR on the visible text and matching it to the known content of the page. We used the Tesseract OCR library[23] for text detection and recognition of each preprocessed frame, then compared the results to the pre-computed document text. For efficiency, we tracked character motion between frames and only performed OCR when sufficient motion had occurred or when the system was unable to reliably estimate the current location (allowing us to achieve an average processing rate of 20–30 fps). Because the camera's field of view was large enough to encompass multiple partial words across 2-3 lines of text (Figure 4.10c), the system did not generally encounter difficulty distinguishing the locations of repeated words. The likelihood of this potential problem was further reduced using recent location estimates and the motion of the user's finger to resolve conflicts. We tracked the current line of text using the camera's estimated motion and the known content of the page, and only provided text-to-speech feedback when the user advanced on the current line. In order to provide a smooth reading experience, it was not possible to skip or repeat words. Although this enforced sequential reading of the text, it mitigated several potential sources of confusion that would have arisen had we allowed rereading

[23] https://code.google.com/p/tesseract-ocr/

or moving between lines. The software detected that the user had reached the start or end of a line or paragraph using the known content of the page, and provided the same audio cues as in Study I. Also, as with Study I, the speed of the text-to-speech feedback was adjusted to match the user's finger speed.

The test apparatus for the second part of the study consisted of the KNFB Reader iOS application running on an iPhone 5S with the VoiceOver feature enabled. To take a picture, users tapped on the left side of the screen to select the "Take Picture" button, and then double-tapped the button to capture an image. The software played a shutter sound to inform the user that the picture was captured successfully, and then immediately began reading any recognized text.

Procedure. These exploratory study sessions lasted 1–2 hours. The participant first used HandSight with his or her preferred directional guidance method from Study I. As with Study 1, training and testing documents were selected from the *Iowa Test of Basic Skills*. For training, the experimenter first re-introduced exploration mode and asked the participant to explore a plain document for up to three minutes. Participants were directed to count the number of paragraphs and columns, and to note the size and position of the margins. The experimenter then re-introduced reading mode's audio cues and directional guidance, and helped the participant read the training document, providing verbal or physical guidance if necessary. The training tasks lasted 10–15 minutes. After training, participants explored and read one single-column test document, with the experimenter providing verbal assistance only if the participant was unable to proceed. Afterward, participants answered questions about the layout of the

document, three multiple-choice questions to judge comprehension, and subjective questions about the experience. We did not use a magazine-style document because the HandSight camera prototype does not currently support two-column documents.

Following the use of HandSight, the experimenter introduced KNFB Reader iOS: how to position the phone's camera over a page, take a picture, and listen to the recognized text. Although the KNFB Reader iOS application included a spoken field of view report to assist with framing a document, we did not evaluate this feature due to time constraints and because it was not the focus of this study. Participants were allowed to repeat this process up to three times with a single-column training document, with verbal or physical guidance as needed. This training task lasted 10–15 minutes. Participants then read two test documents unassisted: a single-column document (from the Iowa Test of Basic Skills) and a two-column magazine document (from USA Today) similar to those read in Study I but without images. KNFB Reader iOS advertises support for multicolumn formats, and the procedure for capturing and reading the two types of document was identical. If the participant was unsatisfied with the reading result, they were allowed one additional attempt per document. Participants answered multiple-choice comprehension questions after the single-column document and summarized the content of the two-column document. Finally, participants reported on their experience using the application. See Appendix B for the full text of the subjective questionnaires for both HandSight and KNFB Reader iOS.

Participant Identifier	P10	P11	P12	P19	Mean
Guidance Type	Audio	Haptic	Audio	Haptic	N/A
Start of Text	5	2	5	1	3.3
Start of Line	5	2	2	2	2.8
End of Line	5	5	5	5	5.0
End of Paragraph	3	5	5	5	4.5
Line Tracing	2	2	3	2	2.3
Understanding Cues	5	5	5	3	4.5
Reading and Understanding	3	3	5	4	3.8
Mean Ease of Use Rating	4.0	3.4	4.3	3.1	3.7
Average Reading Speed per Line (wpm)	18.4 (*SD*=5.5)	56.6 (*SD*=16.4)	60.2 (*SD*=11.1)	44.9 (*SD*=17.1)	45.0
Average Line Finding Time per Line (s)	30.5 (*SD*=24.4)	8.8 (*SD*=5.6)	7.3 (*SD*=5.0)	18.0 (*SD*=12.7)	16.15
Time to Read Full Document (s)	1493	469	409	717	772
Comprehension Questions Score	2/3	3/3	3/3	3/3	2.75/3

Table 4.5: Top: Ease of use responses while using the HandSight prototype. Responses range from *1 - very difficult* to *5 - very easy*. Bottom: Performance metrics from the HandSight reading task. The document for this task consisted of 282.6 words (normalized to 5-character length) across 17 lines.

4.2.2 Findings

Our findings are exploratory, including general observations about how participants approached the reading tasks, and subjective responses to both our proof-of-concept implementation and KNFB Reader iOS. While the focus is on qualitatively describing experiences with the technologies, we include performance statistics such as reading speed, line finding time, and number of skipped words.

Overall Experience. All four participants completed the reading tasks, but with varying levels of success (Table 4.5). P10 read slowly and required frequent verbal and physical intervention by the tester to adjust hand position and answer questions about the directional audio cues. P11 and P19 read more quickly, and only needed infrequent verbal reminders (P11 was reminded once about hand position, and P19 was reminded once about hand position and the procedure for finding the start of a line). P12, who was very comfortable with both braille and screen readers, read the fastest, at 60.2

wpm, and did not require any assistance. Only P10 failed to answer all three comprehension questions correctly, likely due to decreased attention to the content while struggling to complete the task.

Comments were similarly mixed. P19 was enthusiastic about the concept, stating: *"I'm very pleased and excited about the system. I think it could make a great difference in my life."* P12 was more critical, finding the approach to be slower than expected: *"It seems like a lot of effort for reading text."* P10, P11, and P19 were all learning to read braille at the time of the study, and P11 and P19 found the reading experience using HandSight to be easier than braille for reasons similar to those expressed in Study I (*e.g.,* lower learning curve, less to remember). P10 stated that braille and finger-reading were both difficult at times, requiring too much concentration to read quickly or fully comprehend the text. P12, who had the most braille experience, found HandSight to be "much worse" than braille and "somewhat worse" than other technologies for reading printed documents. In addition to commenting on the ease of following a line of braille text due to the tactile feel of the dots and the lack of layout issues such as multiple columns, P12 said that he typically scans printed documents to read on his computer or mobile device, an approach he finds faster compared to HandSight and one that does not require the use of both hands.

Cognitive Load. Although they were able to complete the reading task, all participants expressed concern about the level of concentration required to interpret the directional guidance and other audio cues while listening to synthesized speech. P11, for example, commented on the difficulty of remembering how to map the haptic

guidance to up/down movement: *"it gets you a little confused sometimes, especially if you was [sic] into reading the story and you forget which one was the vibration for moving up and which one was for moving to the bottom."* P11 also commented on the focus and practice required, concluding that it would be difficult to use, *"if you're tired, if you've had a long day."* More practice with the device may address some of these issues, though interaction design changes are also likely needed (*e.g.*, more intuitive and responsive directional cues to reduce required concentration on line tracing task, efficient rereading to enhance comprehension).

Technical limitations with the prototype may have exacerbated cognitive load issues. Although our algorithms ran at approximately 30 fps on average, they tended to run more slowly after rapid finger movements. This limitation caused a noticeable lag at times, which P11 and P19 reported required more concentration. P19, for example, commented that after the start-of-line audio cue there was sometimes a delay before the speech began, causing problems: "*I wasn't getting that in my head to just wait for the delay. I started moving my finger"*.

Physical Design. Three participants identified limitations with the prototype's physical design. The primary issue stemmed from the camera placement: for the text to be an appropriate size and orientation within the camera's field of view, participants' hands needed to be held at a specific angle. Although the camera's placement on the finger was adjusted at the start of each study session, it could not easily be readjusted. Participants thus had to hold their hand at nearly the same angle throughout the study. Two participants reported that this position was too uncomfortable for extended use,

(a) (b) (c) (d) (e)

Figure 4.11: Examples of situations where HandSight was unable to provide feedback. All images have been preprocessed to emphasize text and highlight baselines for the current line. (a), (b) Not enough text is visible in the margins to provide directional guidance. (c) The camera position changed after calibration and is too far from the page to reliably recognize text. (d) The camera is moving too quickly, blurring the text and reducing the frame rate of the recognition algorithms. (e) The user's middle finger is in the camera's field of view, preventing correct segmentation of the lines of text.

suggesting that the physical design will need to improve in future versions and/or the camera location should be easily adjustable. We also identified the need for feedback when the system loses its position in the text or is unable to recognize visible text in reading mode (Figure 4.11 shows examples); this occurred when the hand position changed too much or, more commonly, when the participant moved into the upper or lower margins of the document.

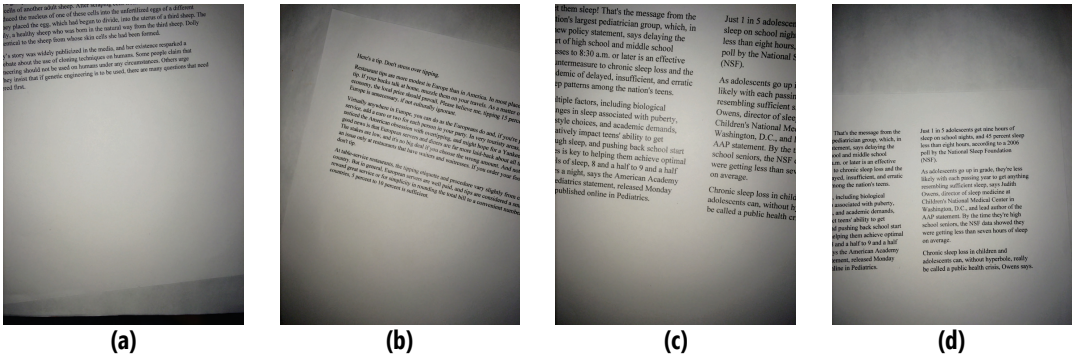
HandSight vs. KNFB Reader iOS. While the study did not offer a controlled comparison of HandSight and KNFB Reader iOS, we can draw preliminary conclusions about tradeoffs between the two. Even without KNFB Reader iOS's document-framing guidance enabled, participants unanimously preferred it to HandSight, with three participants rating it as *5 – much better* and one as *4 – somewhat better*. The primary reason was the fluidity of the reading experience after capturing an image with KNFB—the application read the full document quickly and participants were able to concentrate solely on the content of the passage. For example: *"It just did it all for you, that way you just listen to what it's saying and then take in the details"*

(P11). P12, who had previous experience with KNFB, also stated: *"I like that the text is immediately available to use for other purposes […] I can go back and review the text letter by letter if I need to."* The average reading time was only 187 seconds for the first document and 146 seconds for the second, even with the two attempts that participants were allowed, as compared to an average time of 772 seconds to complete the reading task with HandSight (Tables 4.5 and 4.6).

Although participants preferred KNFB Reader iOS overall, the process of capturing an image was not always straightforward without the document-framing guidance. P11, for example, said: *"It was easy to read it once you got it right, but it was difficult to center [the camera] in order to get the whole text"* (P11); see Figure 4.12 for examples of images captured by participants during this study. With the

Participant Identifier	P10	P11	P12	P19	Mean
Document 1: Number of Attempts	2	2	1	2	N/A
Document 2: Number of Attempts	2	2	1	2	N/A
Document 1: Total Time (s)	230	198	93	225	187
Document 2: Total Time (s)	138	219	89	137	146
Document 1: Text Lost (%)	29.7%	48.6%	0.4%	10.4%	22.3%
Document 2: Text Lost (%)	51.5%	0.0%	0.0%	51.5%	25.8%
Comprehension Questions Score	1/3	3/3	3/3	3/3	2.5/3

Table 4.6: Performance metrics from the KNFB Reader iOS reading tasks. The amount of text lost includes both cropped and misrecognized words, and the percentages indicate the best performance out of the two attempts participants were allowed for each document.



(a) (b) (c) (d)

Figure 4.12: Examples of cases where the KNFB Reader iOS application failed to fully capture the content of a document due to partial visibility or excessive rotation.

exception of P12, all participants required a second attempt to capture each document, and even with a second attempt part of the document was frequently omitted. Although accuracy varied across participants and attempts, approximately one quarter of the documents' content was missed on average (see Table 4.6).

Document comprehension appeared to be similar to reading with HandSight, with three participants answering all comprehension questions correctly (Tables 4.5 and 4.6). However, even when participants were able to understand the main points of a document, the reading experience was not always smooth due to missing text: *"It's not always easy to know if I have the entire page. That was a problem with the first test document. While it was still understandable, I clearly lost some of the text"* (P12).

Summary of Study II Findings. While this study was not meant to be a controlled comparison of finger-based reading versus a mobile scanner, it offers some guidance for future studies. HandSight provided more immediate access to text content than KNFB Reader iOS but was much slower and was perceived as requiring a greater level of concentration. Once the document was successfully scanned, KNFB Reader iOS offered a faster and smoother reading experience and was preferred by all participants. HandSight provided additional information about the spatial layout of documents, but further investigation is needed to determine the impact that may have on document understanding.

4.3 Discussion

Our findings highlight tradeoffs between haptic and audio directional guidance for finger-based reading. We also reflect on the feasibility of finger-based reading compared to existing methods, and outline ideas for iteration on HandSight's design.

4.3.1 Audio versus Haptic Directional Guidance

For blind users, effective finger guidance is critical for line-by-line reading, and therefore directly impacts the feasibility of the finger-based reading approach. Although there were few statistically significant differences between audio and haptic finger guidance in Study I, some tradeoffs emerged. For the magazine documents, audio guidance resulted in significantly more accurate line tracing than haptic guidance. The exact cause is unclear. To scaffold participants in learning how to do finger-based reading, we always presented plain documents before the more complex magazine documents. That audio was more accurate than haptic for the magazine document thus suggests that haptic may have a steeper learning curve, participants have become desensitized to the haptic vibration over time, or that haptic is somehow not as effective with complex document layouts.

In terms of subjective responses, our findings reflect the conflicting results seen in prior work [189,190,199]. Out of 19 participants, 11 preferred haptic, 7 preferred audio, and one was undecided. One downside of the audio guidance is that it occupies the same channel as the speech output, which made it difficult for some participants to concentrate on the text-to-speech synthesis. Twelve participants, half of whom even

preferred audio guidance, commented on this issue. For haptic guidance, the potential issue of desensitization or numbness arose even in this short study, suggesting that a longer-term evaluation will be important.

We also encountered disagreement over how audio and haptic cues should map to up/down direction, which could have impacted results. The mappings used in our studies were the result of pilot testing and our experiences in [199]. For audio, we used high pitch to indicate that the finger should move up and low pitch for down. For haptic, the vibration motor on the underside of the finger indicated downward movement, while the top vibration motor indicated upward movement (in essence, pulling the finger). While the majority of Study I participants were satisfied with these mappings, 4 felt audio should be reversed and 3 felt haptic should be reversed. More work is needed to identify which mapping is best for both audio and haptic, or whether additional training time would mitigate the issue. Ultimately, this setting may need to be user-configurable. Future work should also investigate alternative feedback approaches (*e.g.,* the pitch of the speech synthesis could provide directional guidance).

4.3.2 Feasibility of a Finger-Based Reading Approach

We had expected a finger-based reading approach such as HandSight or Shilkrot *et al.'s* FingerReader [189,190] to offer many advantages over mobile-based scanners for reading printed text: access to spatial layout information, direct as opposed to sequential access to text on the page, reduced camera framing issues, and, compared to crowdsourced approaches *(e.g.*, [11], BeMyEyes), real-time OCR. However, while we

observed some of these advantages, important concerns also arose. Here, we reflect on the feasibility of finger-based reading, incorporating ideas for future work throughout.

Document Layout and Spatial Awareness. A primary motivation for investigating HandSight's finger-based reading approach was to provide users direct access to spatial layout information. Our exploration mode provided audio cues to indicate text, pictures, or white space beneath the user's finger. While four participants in Study I commented positively and unprompted on this information, one participant was strongly against the idea, feeling that software that could automatically process a document's layout to extract content would be preferable in most situations. The difficulties encountered by the participant who was removed from our dataset in Study I also highlight an unexpected but important potential for confusion: some users, particularly those who are congenitally blind, may have an inaccurate or incomplete understanding of basic document structures (*e.g.,* columns, margins) simply because they have never encountered them. In that participant's case, he was not familiar with the notion of columns, which led to confusion. Future work should explore the relationship between a user's spatial abilities and their proficiency in exploring a document or responding to finger guidance.

While exploration mode helped participants understand a document's layout (*e.g.,* number of images), distinguishing a gap between paragraphs versus columns was particularly challenging. Both types of gaps were indicated by white space, but participants were frequently unable to determine whether the white space occurred between two paragraphs or between two columns. We intended for paragraph gaps to

be distinct from column gaps by the direction in which the finger is moving—vertically for paragraphs or horizontally for columns. However, without sight, many participants tended to move their finger more diagonally, drifting accidentally between paragraphs and columns. This challenge could be addressed by designing cues to identify the horizontal and vertical edges of a block of text.

Finally, we did not evaluate the potential utility of layout information for blind readers. And, arguably, for half the documents we used (the plain text documents), spatial information offered little benefit. The finger-based reading approach may be more beneficial for other types of documents, particularly those with inherent spatial characteristics such as maps or graphs.

Cognitive Load and Physical Effort. Our studies indicate that line-by-line reading incurs high mental and physical effort. The reader must simultaneously attend to directional guidance, document events (*e.g.,* start and end of line), and the synthesized speech content. Study II, in particular, highlighted the increased concentration and physical dexterity required to use HandSight compared to KNFB Reader iOS. This issue of physical effort confirms previous findings from a much smaller study (3 participants) [190]. With more practice, HandSight should not require as much effort to use, and, if the technology provides enough benefit, the need for this practice is not necessarily a barrier to adoption—braille and the Optacon [76,138], for example, require extensive practice. However, a multi-session study would be needed to assess just how much practice is needed and how efficiently experienced users can read with a finger-based approach.

Camera Placement. Whether they use crowdsourcing or automated OCR, both mobile document scanning approaches (*e.g.*, KNFB Reader iOS) and body-mounted solutions (*e.g.,* OrCam) require a global image of the document, properly aligned and in focus within the camera's field of view. All participants in Study II reported at least some difficulty with this type of image capture using KNFB Reader iOS, but we had not introduced them to KNFB's document-framing feature. That feature, along with findings from blind photography research (*e.g.* [36,86,213]), should help overcome the issue. The global image captured by KNFB also allowed for more fluent text-to-speech than with HandSight, which participants valued. At times, however, our own use of KNFB Reader iOS and observations of participants showed that this fluency can provide a false sense of confidence. That is, it is not always clear from the speech output if a part of the document is missing or the application has parsed and played text blocks in the wrong order.

HandSight's finger-mounted camera and direct control over text scanning and speech playback may overcome these issues to some extent, but Study II showed it also introduces new camera placement challenges. For example, participants frequently encountered difficulties tracing lines near the upper and lower margins because the system could not provide directional guidance when no text was visible.

Future work could explore hybrid methods that may combine a body- or head-mounted camera with a finger-mounted one, potentially overcoming the weaknesses of each and supporting a wider range of reading situations. A body-mounted camera could capture complete documents and allow for efficient, fluent reading using a screen-

reader interface and relative exploration of content (*e.g.*, swipes). At the same time, the finger-camera interface could provide knowledge about the document layout, acting as a cursor to quickly search through the content or provide contextual information. It would be useful to compare how well access to both types of interaction works compared to only the global, relative interaction or the finger-based interaction.

Physical Design and Social Acceptability. Physical design and social acceptability influence the adoption of wearable technology [169,175,176]. While our early HandSight prototype is bulky, future versions could be substantially reduced in size since the underlying technology (*i.e.,* the endoscopic camera) is extremely small. Still, whether blind users are interested in wearing a finger-mounted device for accessibility is an open question. Social acceptability could also change how users feel about the haptic line guidance compared to the audio guidance in practice. The majority of users preferred haptic guidance in Study I, but even in future iterations of the physical design, the haptic vibration motors would likely add bulk compared to audio alone. These issues are not unique to HandSight, and the question of where users will feel most comfortable having a camera mounted on their body (if at all) should be explored in future work.

Target Users. While our prototype was designed to support totally blind users, the question of who may benefit most from a finger-based reading approach remains open. Low vision users, for example, may benefit from the direct access and physical gestures that a finger-based reading approach provides, without finding the line tracing as time consuming as for a totally blind reader. We recruited one low vision pilot

participant, who experienced no difficulties with describing the layout of the document or with line finding and line tracing. The device could then act as a more portable alternative to closed-circuit television (CCTV) magnifiers, automatically processing the words and providing additional information about the text upon request (*e.g.*, font, spelling). Further investigation, however, is needed to explore this possibility and how it is received compared to commonly used magnifiers.

4.3.3 Design Iteration

In addition to the future work mentioned above—such as investigating the utility of spatial layout information, conducting a longer-term study, and evaluating HandSight with low-vision users—our findings lead to several design revisions that may improve blind users' experience with HandSight.

We designed the speech interface to adapt to the user's finger speed to easily control the rate of feedback. Some participants liked this feature, but others found it to be uneven when compared to the continuous speech feedback of screen reader software, noting that it was difficult to identify the end of a sentence. More fluid speech feedback and additional audio cues to mark punctuation could ease the reading experience.

An important observation from Study II is that a finger-based reading device should provide an easy way of determining when text is no longer contained within the camera's field of view. Participants occasionally confused situations where the system could not provide guidance (not enough text in the frame) with being correctly centered over the current line and not receiving directional feedback. To address this issue and

provide users with more information while reading, document exploration and reading modes could be integrated. In doing so, however, we must take care not to further increase cognitive load and distract from the content of the text.

To reduce the image capture issues seen in Study II, another possibility is to redesign the physical prototype to either move the sensor farther away from the text (as with FingerReader [189,190], which is on the upper part of the finger) or to use a wide-angle lens. These options could expand the camera's field of view, for example, allowing users to drift farther away from a line before the text is lost.

Finally, our prototypes only allowed users to continue reading forward and did not support backtracking, rereading, or jumping to an arbitrary location in the text. Study I focused on sequential line-tracing guidance, but it would be interesting to implement and evaluate these additional reading actions.

4.3.4 Limitations

Using an iPad rather than a physical prototype to compare haptic and audio line guidance in Study I was a conscious study design choice, allowing us to bypass technical challenges in implementing a real-time prototype and to focus on the user experience and collect precise line traces. A limitation of this choice, however, is that the experience of reading with a physical prototype and paper may be different. As well, the font size and document layout for Study I were constrained to two specific formats, which do not fully represent the variety of real-world documents that users may encounter. In Study II, an important limitation of the proof-of-concept prototype

is that we assumed that the content of the page was known prior to beginning reading and constrained the system to allow participants to read text sequentially from left to right and top to bottom. These choices simplified how the system provided finger guidance—it only needed to estimate the finger location on the page and provide upward or downward guidance to return to the last known line. However, these artificial limitations also disregarded some of the potential advantages of a finger-based reading approach, such as re-reading or jumping to arbitrary locations. We also asked participants to hold their hand in a specific position for Study II, constraining their natural behavior when using a device such as ours. Study II was not meant to offer a controlled comparison of KNFB Reader iOS and HandSight, but limitations even for gathering exploratory feedback include that we did not evaluate the document framing feedback of KNFB Reader iOS, and that only one participant had previous experience with KNFB Reader iOS (all 4 had used finger-based reading). While a more controlled comparison is thus needed, it is important to note that participants still identified many strengths of KNFB Reader iOS. Finally, while we focused on blind readers, it would be interesting to expand the evaluation of finger-based reading to users with a wider range of visual abilities.

4.4 Summary

We conducted an in-depth study with 19 blind participants comparing audio and haptic cues for directional guidance to support finger-based reading. Our findings showed similar performance and user preference between the two types of guidance, although

audio resulted in significantly more accurate line tracing for some tasks. Subjective feedback was split but suggests that haptic guidance may be slightly preferred. In addition, our findings highlighted general strengths and weaknesses of a finger-based reading approach, such as improved understanding of a document's layout and the difficulty encountered by blind users in accurately tracing a line of text with a finger. In follow-up sessions where 4 of the participants used a proof-of-concept finger-reading prototype as well as KNFB Reader iOS, the mobile scanner was seen as offering a more fluent reading experience. Ultimately, a finger-based reading approach may be best suited to material that is inherently spatial, such as maps or graphs, whereas existing applications that capture a global image of the document for text-to-speech (*e.g.*, KNFB Reader iOS) may be preferred for text-heavy material. Future work should investigate this possibility, as well as assess the potential of finger-based reading for low-vision users, for whom precise directional finger guidance may not be necessary.

Chapter 5: Augmented Reality Magnification for Low Vision Users

Our work thus far has focused on helping users who are totally blind to read printed materials, but a similar device could also benefit low vision users without the need for complex audio and haptic guidance. Furthermore, the ability to provide visual feedback alongside the other channels opens up a new realm of design possibilities. This chapter explores recent advancements in augmented reality (AR), which have the potential to increase the quality of life for people with visual impairments. For low vision users, head-mounted displays (HMDs) that enhance existing visual capabilities are particularly promising. For example, ForeSee [235] used an Oculus Rift VR headset with an attached camera to magnify and enhance text content, and other researchers used Google Glass to enhance edges within the wearer's field of view [82] or display magnified content from a smartphone screen [170]. Several commercial HMDs (*e.g.*, eSight [238], NuEyes [239], IrisVision [240]) display magnified video captured from a head-mounted camera, and provide image enhancement features such as contrast adjustment. A recent study investigating the use of one of these systems (eSight) was generally positive, showing the impact HMDs can make in users' lives [237].

While these systems have begun to explore how HMDs and wearable cameras can be used to augment visual perception, they are limited to enhancing and/or magnifying the 2D image from a video camera. In contrast, the classical definition of

---

This chapter contains work scheduled to be published in the proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2018) [200].

AR integrates 3D virtual objects into the 3D physical environment [10], which would allow for new visual enhancement possibilities that are better integrated with the user's real-world tasks. For example, a magnified view of an object can be rendered directly on top of the real object, fixed to a desk near the user's primary work focus, or "projected" on a nearby wall. Off-the-shelf technologies such as the *Microsoft HoloLens* [241], an optical see-through display, are beginning to have the capability to support these types of 3D AR designs.

To investigate the design possibilities for AR magnification tools enabled by registering virtual content in real 3D space, we conducted a series of iterative design sessions with seven low-vision participants. We developed initial prototype designs on a Microsoft HoloLens, which we presented to participants to solicit feedback and open-ended ideas about future wearable magnification aids. Our designs explored several different virtual display options (*e.g.,* affixed to real objects *vs.* moving with pointing finger), image acquisition approaches (head-mounted, finger-mounted, or smartphone),

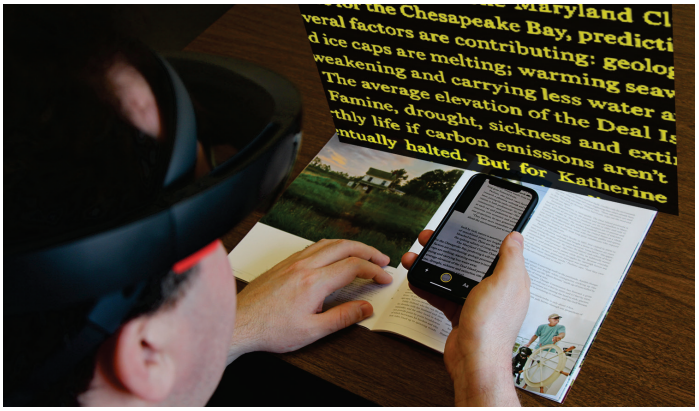


Figure 5.1. Prototype AR Magnification system using a transparent HMD (the Microsoft HoloLens) and a handheld smartphone (iPhone X) as a camera and input device.

and interaction techniques (*e.g.,* Figure 5.1). The designs were updated between sessions based on participant feedback as well as our own observations.

Overall, participants liked the concept of a wearable AR magnification aid, especially the natural reading experience and ability to multitask that the projected 3D renderings enabled. At the same time, our system presented some difficulties compared to participants' existing magnification aids. We discuss these issues along with potential solutions and design implications.

Our contributions include: (i) an exploration of the design space for augmented reality magnification; (ii) proof-of-concept implementations evaluated and refined through iterative design with low vision users; and (iii) common themes and recommendations that should inform the design of future AR vision enhancement aids for low vision users.

5.1 A Design Space for Magnification Aids

To inform the design of an AR magnification aid for low vision users, we first outline our goals and key design dimensions for mobile and/or wearable magnification aids.

5.1.1 Design Goals

Informed by prior work, existing commercial systems, and our own experience working with initial AR prototypes, we formulated the following design goals for our study:

- **Augment rather than replace.** Whenever possible, avoid interfering with the user's existing vision capabilities. Provide enhanced content alongside the real world with easy controls to hide or reposition the digital information as needed.
- **Leverage augmented reality.** Go beyond the static 2D displays provided by existing systems and explore applications for persistent digital content overlaid in 3D onto the physical world.
- **Prioritize customization and flexibility.** To support a wide range of vision levels and different situations, the ability to customize how the enhanced content functions and appears is crucial [92].

5.1.2 Design Dimensions

To achieve these goals, we considered several design dimensions in addition to virtual display position, our primary dimension of interest:

- **Virtual display position.** The ability to anchor virtual content to a physical location in 3D space enables several possible virtual display designs. Specifically, we explore four positions. The first, simplest position is a fixed heads-up display that moves with the user's head to always stay within their field of vision. The second position is a stationary display attached to a location in the physical world, which maintains its position as the user moves. The third option is a dynamic display that acts as a magnifying glass and follows the user's hand or other moveable object (*e.g.,* a ring or smartphone). Finally, the

fourth position projects an image directly onto the physical object that is being enhanced (*e.g.,* a magnified view shown atop a document).

- **Content capture.** To capture video for processing and display, possible camera locations include head-worn (*e.g.,* [159,235,238–240]), hand-held (*e.g.,* [242–244]), and finger or wrist-worn (*e.g.,* ring or smartwatch; [190,194,197]). We explore these options and discuss the advantages and disadvantages of each.

- **Image enhancements.** To support a range of vision levels, important enhancements include magnification, changes to brightness and contrast, binary thresholding, and color alterations (*e.g.,* as described in [235]). Although not the focus of our study, optical character recognition could also be useful, either to read text aloud or to visually enhance the detected text by increasing the resolution or replacing fonts.

- **Physical HMD.** Several display types have been explored previously. However, an optical see-through display and 3D sensing capabilities are needed to achieve our design goals, making the Microsoft HoloLens the obvious choice at the time this research was conducted. The HoloLens allowed us to rapidly prototype and iterate on AR designs; however, we fully expect that future HMDs for AR will be more streamlined, lightweight, and portable (*e.g.,* integrated into traditional glasses).

- **User input.** To support our goal of customizability and flexibility, the AR system needs to provide intuitive and easy-to-use interactive controls. A few

options include physical controls on the device or a separate remote (*e.g.,* eSight [238], Glass [56]), gaze tracking, midair gestures, and voice commands (*e.g.,* OrCam [159], HoloLens [241]), or 3D tracking of a physical object (*e.g.,* Oculus Rift controller [245]). We explore a few of these options to see how well they work for low vision users and in different situations.

5.2 Iterative Design of a Prototype System

To explore these design dimensions and evaluate which designs and features would work best for low vision users, we conducted a series of iterative design sessions. These sessions were structured to elicit general feedback and open-ended design ideas from participants, drawing on elements of cooperative [58] and participatory [185] design methodologies. Based on ideas from existing magnification aids, knowledge of available hardware capabilities, and our own experience working with low vision users, we developed an initial prototype system that implemented several user interface designs. We then asked participants to use the system and provide feedback, refining our design over nine design sessions with seven unique participants (two participants returned for a second session). While we modified the system between sessions to fix issues and make minor improvements, for ease of presentation, we group our prototypes into three basic designs based on the broad design elements, components, and user interactions.

5.2.1 Initial Investigation: HoloLens Only

Our initial design used only the HoloLens headset. As mentioned earlier, the HoloLens includes an optical see-through display on which translucent virtual objects ("holograms") can be overlaid in real 3D space. The estimated field of view is 30◦ × 17.5◦ with 2500 light points per radian. Microsoft's APIs include motion tracking and 3D scene analysis functions that allow developers to anchor digital content to a physical location in the world so that it will remain stationary as the user moves. The APIs also support input using midair gestures, the direction the user's head is pointing, and voice.

This initial prototype used the HoloLens's built-in camera to capture images of what the user looked at and provided two modes for displaying an enhanced version of those images: fixed 2D and fixed 3D. While we describe these display modes in more detail in the next section, the fixed 2D display moved with the user's head to always remain within view while the fixed 3D mode was anchored to a surface in the physical world. Users could toggle between modes using voice commands and two image enhancement options were provided: binary thresholding (black text on a white background) and color inversion.

While this initial design was functional, internal testing revealed that the HoloLens' built-in camera resolution was simply too low to be useful when magnified. Additionally, requiring users to turn their head to look at desired content for magnification was uncomfortable, and the voice commands were cumbersome and imprecise. We used these observations to develop the next iteration of our prototype, which was the first to be tested with low vision participants.

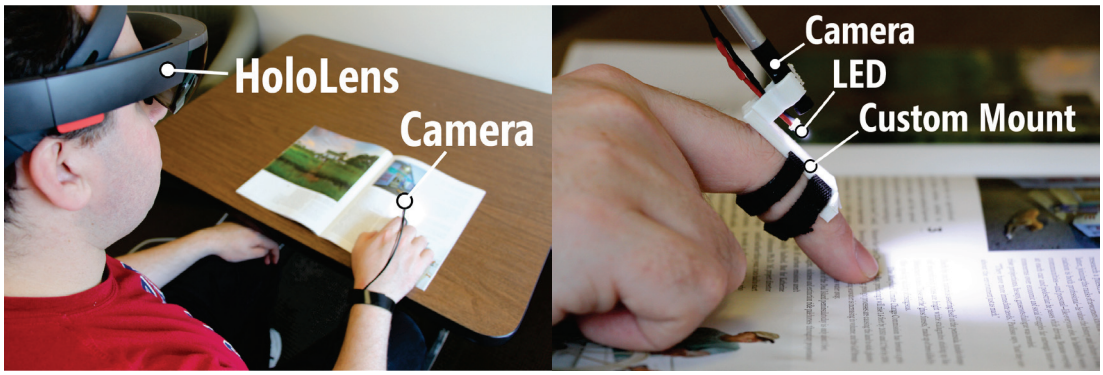


Figure 5.2: First AR magnification prototype system design: (a) full system with the HoloLens, (b) close-up of the finger-worn camera

5.2.2 Prototype 1: HoloLens and Finger-Worn Camera

To address the issues observed in our initial investigations and to expand on our design, we added an external camera, implemented two additional display modes and more customization options, and replaced the voice commands with a virtual menu controlled using midair gestures. We then conducted design sessions with three participants, making minor changes to the system between sessions based on feedback (*e.g.,* modifying the perceptual distance at which the AR displays were drawn, simplifying and polishing user input).

Implementation Details

Below we describe the prototype's components and physical design, the four display modes, and the user interactions.

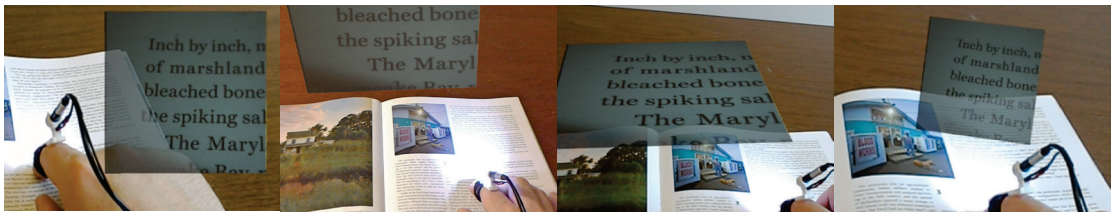
Hardware and Physical Design. This prototype used an external camera[24] mounted on the user' finger using a custom 3D-printed ring with Velcro straps and an

[24] Awaiba NanEye GS Idule Module, 640×640px, 30° FoV

LED to provide consistent lighting (Figure 5.2). The camera provided a close-up view of the target content and allowed the user to read without needing to frame the text within the head-worn camera's field of view. As discussed earlier, similar wearable cameras have been used for other assistive devices [146,190,198,203], albeit primarily for people with more severe visual impairments. We used a laptop computer to capture and process images from the camera, which we streamed wirelessly in real-time to the HoloLens for display.

Virtual Display. To elicit feedback on a range of AR display designs, we implemented four options for displaying the enhanced view from the camera, including the two explored in the initial prototype (Figure 5.3 and video figure):

- *Fixed 2D:* This design displayed the image at a fixed location relative to the user's head, with the display within the user's view at all times. This design is similar to past work using HMDs for visual enhancement (*e.g.,* [82,235,238]), and it is possible to implement on more basic HMDs such as Glass.
- *Fixed 3D Vertical:* This design allowed the user to place the enhanced view from the video camera at a fixed position in the physical world. The display was



(a) Fixed 2D (b) Fixed 3D (Vertical) (c) Fixed 3D (Horizontal) (d) Finger Tracked

Figure 5.3: Prototype 1 provided four virtual display modes, which could be customized (position, size, zoom) using midair gestures. See the accompanying video figure for a demonstration: https://youtu.be/i0IDbHGir-8.

oriented vertically, allowing for placement on a vertical surface like a wall. It remained at the fixed location as the user turned their head or moved around.

- *Fixed 3D Horizontal:* This design was similar to Fixed 3D Vertical but was oriented horizontally so it could be placed on a tabletop or other flat surface.
- *Finger Tracked:* In this design, the display was oriented vertically above the user's finger and moved along with the user's hand like a magnifying glass. Because the HoloLens APIs could not detect the user's hand when touching a page, we used a simple image processing technique to detect the bright LED from the finger-mounted camera and position the display near it.

As with the previous prototype, the system provided image binarization and color inversion features, which participants could use as desired. The brightness of the LED and the HoloLens display could also be adjusted.

User Interactions. Because voice commands proved too limiting for the variety of customization options we wanted to support, we used the gesture recognition capabilities provided by the HoloLens to allow users to adjust the display's position, size, and zoom level. Users opened the virtual menu using the default HoloLens "air tap" gesture, and selected from three options (position, size, or zoom level) by turning their head to position a cursor atop the desired item and then performing another air tap gesture. To move or adjust the display, users performed a "manipulation" gesture, first lowering their index finger, moving their hand in any direction, then raising their finger again once satisfied.

The first two VI participants found these interactions to be difficult, so we reduced the number of menu items (leaving only position and size) and added remote controls to allow us adjust settings as directed by the participant if needed.

Design Sessions

Three participants used our first prototype, comparing the display modes and discussing the overall idea of augmented reality vision enhancement.

Participants. We recruited three participants who used a CCTV or other magnification aid in their daily lives (two male, one female, age range 28–54). The cause and severity of participants' visual impairments varied (Table 5.1).

Methods. After an open-ended interview to collect demographic information and participants' experience with magnification aids, we introduced our system and demonstrated its features. Participants then used each of the four display modes in a partially counterbalanced order (using 3 out of 4 orders from a balanced Latin square) to read text on a variety of objects, including simple printed documents as well as mail, a pill bottle, and a box of cereal. After each mode, participants provided feedback on their likes and dislikes for that particular mode, as well as thoughts about the

ID	S1	S2	Age	Gender	Diagnosis	Visual Acuity (self reported)	Visual Field
P1	✓	✓	28	M	LHON	20/400 or 20/450	Limited central vision
P2	✓		46	F	Retinitis pigmentatosa	Low vision (acuity unknown)	Limited
P3	✓	✓	54	M	Optic atrophy (meningitis)	20/200	Full
P4		✓	29	F	Tumor	Low vision (acuity unknown)	None in left, tunnel vision in right
P5		✓	58	M	Cone-rod dystrophy	Light and shapes (acuity unknown)	Limited central vision
P6		✓	33	F	Oculocutaneous albinism	20/400 in good lighting	Full
P7		✓	68	F	High myopia	20/400, better in ideal conditions	Full, but better peripheral vision

Table 5.1: Demographic information for the participants across all co-design session. Columns "S1" and "S2" indicate participation in sessions with prototype 1 and prototype 2, respectively.

customization options. At the end of the study, we asked about experiences using the system and which display modes were most and least preferred, discussed the overall design of the system, and elicited suggestions for improvements and new features. Each session lasted approximately two hours, and participants were compensated $60 for time and travel costs.

Overall Response and Display Modes. The participants each used the system to read the provided materials, with varying levels of success. P1 and P3 reacted positively to the concept of a wearable AR magnifier. P1 commented:

> *"If there was something I could just wear and then be able to see something better, point the camera at it and then have it right there in front of my eyeball then I would use that all the time… You could certainly do many things that you're not able to do by yourself at this point."*

Both P1 and P3 observed advantages to the 3D design elements incorporated into three of the display modes. They considered the two fixed 3D display modes to be more like the reading experience with a CCTV or handheld magnifier than the other two designs, while the dynamic finger tracking design could potentially help to quickly locate a particular location in a document.

Overall, P1 preferred the two fixed 3D designs (either vertical or flat) because they were steadiest and easiest to read, while P3 preferred the fixed 2D design because it was always visible and required the least concentration to use. In contrast to the other two participants, P2 found the reading process too difficult and did not see advantages

to the AR magnification approach, stating that she would prefer to use audio output from a screenreader for most reading tasks.

All three participants disliked the dynamic finger tracking display, primarily due to technical issues with our implementation. This design required participants to turn their head to look directly at their finger, which had two problems: first it was uncomfortable and required additional concentration to keep their finger always within the HoloLens camera's field of view, which interfered with the reading experience. Second, the bright LED reduced the contrast of the virtual display and made it more difficult to read the enhanced text. Interestingly, P1 also found the Fixed 2D display to be difficult to use because its perceptual distance was fixed too far away—we made this setting adjustable for future participants.

Finger-worn Camera. Perceptions of the finger-worn camera were also mixed. The wearable camera allowed for hands-free use, and separation from the display allowed participants to find a comfortable reading position. However, participants disliked the need to move their finger to read (P2) or found it difficult to move from one line to the next for longer passages (P3). The biggest limitation was the small field of view due to the camera's proximity to the page—each image contained only 3–4 lines of text and a few words across. This problem was compounded by the limited field of view of the HoloLens, which when magnified to a readable size sometimes meant that participants could only fit a word or two on the display at a time. All three participants mentioned that their existing magnification aids had a similar problem, but also stated that our system was worse in its current implementation.

The HoloLens Device. Participants' other feedback about the prototype system primarily centered around limitations of our chosen hardware, especially the physical size, weight, and display contrast. Contrast was a source of difficulty for all three participants, to varying degrees. Images displayed on the HoloLens screen are translucent, which meant that text could be difficult to recognize depending on the background imagery. This issue was addressed somewhat by lowering the room lighting or moving the display so that it was positioned over a flat, high-contrast surface (*e.g.,* a white wall or black screen). As mentioned above, the bright LED interfered with reading, so the participants mostly positioned the virtual displays so that they were not looking directly toward it. Even with these measures, P2 was unable to successfully use the system to read because of how the HoloLens display functioned, only able to make out a few scattered words and letters. This finding fits with previous mixed results using optical see-through displays for low vision users [234], and suggests that the HoloLens may work better for some types of visual impairment than others.

User Input. While P1 and P3 were able to use the midair "air tap" gestures to adjust the display, all three participants found the gestures to be cumbersome and difficult to use. We frequently needed to assist with changing settings. Because of these difficulties, participants may not have fully customized the system to meet their specific needs. Additionally, the combination of the slow input and the camera's physical positioning meant that participants could not quickly adjust the magnification level to help with locating the start of a new line or another desired location in a document.

Technology Comparisons. When asked to compare the device with their existing magnification aids, all participants stated that the current version was less convenient, primarily due to limitations with the physical hardware. However, if those issues could be solved, one participant stated:

> *"In comparison to the portable CCTV I have or the full size one, this would be something you could wear and take with you… If you just have a pair of glasses that could essentially do the same thing [as a phone] then I would probably use that even more than my phone." (P1)*

Summary. Two participants reacted positively to the idea of AR magnification and observed potential advantages to our 3D display modes. Hardware and user interface issues—especially the field of view, image contrast, and midair gestures—limited the usability of our prototype, with one participant unable to use the system to read at all. Despite these issues, the overall concept showed promise.

5.2.3 Prototype 2: HoloLens and Smartphone

To address these Session 1 study findings, we redesigned several aspects of our system (detailed below). We then recruited six participants for further design sessions using the updated prototype.

Implementation Details

Below we describe changes to the prototype's components and physical design, display modes, and new user interactions. Figure 5.4 shows the updated prototype in action.

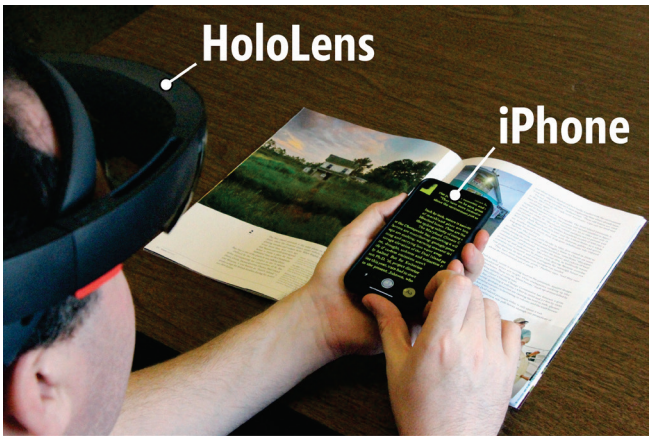
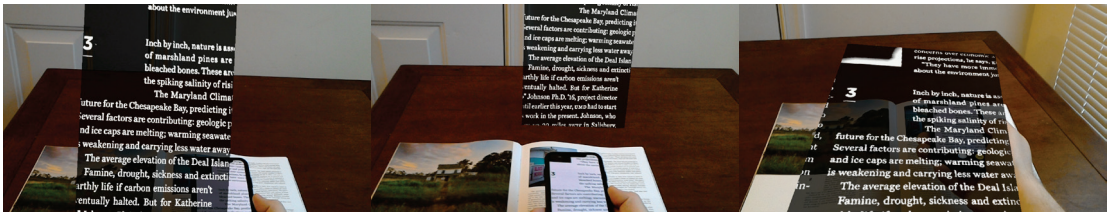


Figure 5.4: Second prototype AR magnification system using the HoloLens and a hand-held iPhone X.

Hardware and Physical Design. Because the finger-worn camera had a narrow field of view and required manual focus, we decided to instead experiment with a handheld smartphone camera (an iPhone X). The smartphone is heavier than the finger-worn camera and does not allow hands-free usage—a feature of the previous design that participants found appealing—but the change allowed us to use a higher-quality camera with built-in processing and wireless communication capabilities. In particular, the camera's high resolution and autofocus allowed users to easily control the amount of text captured by moving the phone toward or away from the page. Users could also adjust the brightness of the phone's camera flash LED to help with contrast. We imagine that a future wearable device (*e.g.,* ring or smartwatch) could incorporate these features as well, if they proved useful for applications like this one.

The use of a smartphone also enabled several new user interactions to control the display settings using the touchscreen and motion sensors, which we discuss below. The phone connected wirelessly with the HoloLens to stream video, touchscreen gestures, and 3D motion data.



(a) Attached to Headset (b) Attached to World (c) Attached to Phone

Figure 5.5: Prototype 2 provided three virtual display modes, which were refined versions of the four included with Prototype 1. See the accompanying video figure for a demonstration.

Virtual Display. Our updated prototype provided three display modes (Figure 5.5; video figure), which were refined versions of the four tested previously. We differentiated the modes by the object to which the display was attached:

- *Attached to Headset.* This display mode was based on the Fixed 2D mode described previously, but with finer-grained control over the relative position and angle of the display. Users could place the display at a location in front of them, and it would move and rotate with them as they turned their head or moved their body, always maintaining the same relative position.

- *Attached to World.* This mode combined the two Fixed 3D designs from the previous study into a single flexible approach that allowed users to position the display in the physical world at any location and angle. As with the earlier designs, the display remained fixed at that physical position as the user moved.

- *Attached to Phone.* This mode functioned similarly to the Finger Tracked design from the previous prototype. It positioned the display atop the smartphone and the display moved as the user moved the phone, acting like a magnifying glass but with an arbitrarily large virtual display. Users could reposition the display—for example, so that it would be vertical while they held

the phone horizontally—but the display would always move to maintain the specified position and rotation relative to the phone.

Aside from the three display designs, we also provided controls for users to adjust the image colors and contrast. As before, we made changes to our designs between sessions. For example, in addition to the black and white color inversion options included for the previous prototype, at the request of P4 we implemented standard white/blue, yellow/blue, grayscale, yellow/black, and red/black filters provided by other digital magnifiers. To address comments from P1 and P3, who were the first to try the new prototype, we also added a "freeze frame" mode which allowed users to press a button to stop the video capture and send a full-resolution photograph to the HoloLens for display. Users could then control the image size and position on the AR display as before, but without needing to hold their phone above the target content while reading.

User Input. Touchscreen controls were used for most input (Figure 5.6), including: double-tapping to open the display mode menu, tapping to select buttons on the screen, pinching to control the size of the virtual display, and sliding to move the display during "freeze frame".

To control the display's 3D position and rotation, we implemented a motion tracking feature using the iPhone's built-in *ARKit* APIs [246]. The API provides functions to track the phone's 3D pose relative to its starting location, which we stream to the HoloLens and use to position and rotate the virtual display. Because the iPhone and HoloLens had different internal 3D reference frames, we manually initialized the

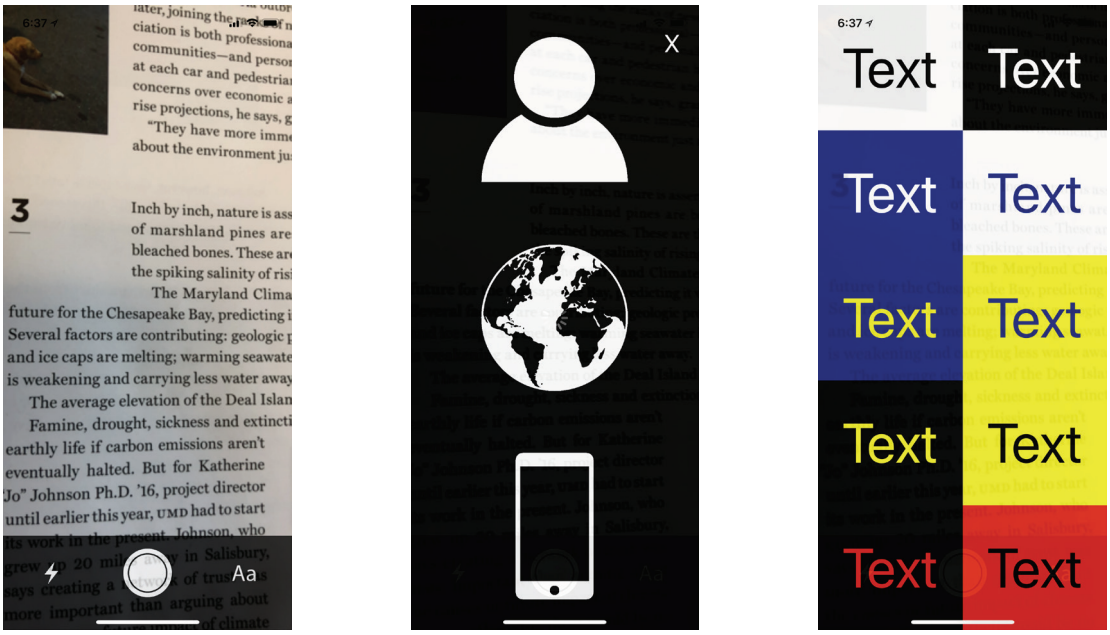


Figure 5.6: Touchscreen controls on the iPhone prototype. Left to right: main screen, display mode menu, text colors menu.

transformation between the two at the start of each session using a simple procedure that required visually positioning the phone atop a virtual representation. This procedure is overly simplistic, and a more robust method will likely be necessary for long-term use. However, it proved sufficiently reliable for the duration of our study.

Users could move the phone to position the virtual screen as desired for each of the three display modes. In both the Attached to Headset and Attached to World modes, users pressed a finger firmly on the screen until the phone vibrated, moved the virtual screen to the desired location (3D position and rotation), then released their finger after they were satisfied with the 3D position and rotation. The interaction was slightly different for the Attached to Phone mode, with users first moving the phone to where they wished the screen to be located, then pressing firmly and moving the phone to where they wished to hold it while reading. After lifting their finger off the touchscreen,

the virtual screen maintained the offset between the initial and final positions as they continued to move the phone.

Design Sessions

Six participants tested our updated prototype, comparing the display modes and providing general feedback as well as open-ended suggestions about their ideal magnification aid.

Participants. We recruited six people with visual impairments (3 male, 3 female, ages 28–68) to participate in design sessions with our updated prototype. P1 and P3 returned from the previous sessions, while four participants had not used our prior prototypes. As with the previous co-design sessions, the cause and severity of the participants' visual impairments varied (Table 5.1) but all participants regularly used some type of magnification aid.

Methods. The user sessions were structured similarly to the previous ones. Participants were introduced to the updated prototype and allowed time to explore the options while reading a simple document. After becoming comfortable with the controls, participants then used each of the three display modes in a fully counterbalanced order to read text on a variety of objects, including simple printed documents, magazine articles, mail, and product labels (*e.g.,* box of cereal, pill bottle). After each mode, participants provided feedback on what they liked and disliked. The session closed with a discussion of participants' overall experience using the system, preferred display modes, and feedback on the system and AR magnification in general, as well as participants' envisioned ideal magnification aid. As with the previous stage,

sessions lasted approximately two hours, and participants were compensated $60 for time and travel.

Overall Response. Participants were in general more successful and positive about the experience of using this prototype than we had observed with the previous version. The iPhone provided higher quality images and better control over the amount of visible text, and the touchscreen and motion controls provided faster and easier control over enhancements and virtual display settings. Participants were better able to experience the augmented reality aspects of our approach, which most participants found promising. One participant was particularly enthusiastic about the Attached to World design, stating that it was:

> *"so much better [than her CCTV], you can go down the whole page and read it. Like if I want to read a book or something to my kids, Mommy doesn't have to go line by line. I can read it and keep the flow going. You can read your whole mail in its entirety without it being on your TV." (P4)*

She felt that our system provided an experience more like what she remembered before her vision loss with advantages to portability and privacy compared to her existing aids, continuing "*It's everything I need as far as being able to read independently"* (P4).

Virtual Display Modes. Participants' display preferences were again mixed, with some participants stating that they liked the flexibility of having multiple designs available and would use different versions depending on the situation. P1 and P3 preferred the Attached to Headset design because they found it easier to focus on the

text with fewer variables to consider. In contrast, P4 found that mode to be too distracting, especially when speaking with someone or otherwise multitasking, and preferred the Attached to World design since it functioned *"like a private, portable CCTV that stays where you want it to stay"* (P4). P5, P6, and P7 saw advantages to all three designs, including the simplicity of the Attached to Headset design, the natural reading experience and ability to multitask with the Attached to World design, and the versatility and intuitive interactions of the Attached to Phone design, especially while away from home (*e.g.,* while shopping). However, all participants found the Attached to Phone design to be more difficult to use than the others for reading longer passages in its current implementation, suggesting that more robust motion tracking and image stabilization are needed to improve the reading experience.

Smartphone Camera. The two participants who had used the previous prototype (P1 and P3) felt that the updated design was an improvement, with a better camera and more usable interactions. However, while the previous design was lightweight and could be used hands-free, the updated design required holding the iPhone steady in midair while reading. All participants found this to be somewhat difficult because of the additional physical effort and shaky image due to unsteady hands. This issue was initially exacerbated by a sometimes slow and uneven frame rate streaming the video from the phone to the HoloLens, which we fixed after the first two sessions, but it also prompted us to add the "freeze frame" feature described above. This feature functioned similarly to existing features on smartphone magnifiers, but with a significantly larger virtual display. Later participants (P4, P6, and P7) liked this

feature and found it to be much easier to use than live video when reading longer passages. The issue of image stability could also be addressed in the future by including a portable mount to help hold the phone steady, by adding optical or digital image stabilization, or by integrating the camera and motion controls into a smaller design (*e.g.,* a smartwatch).

The HoloLens Device. While replacing the finger-worn camera with an iPhone camera addressed one aspect of the limited field of view from the previous design (allowing more text to be captured at once), the narrow window that the HoloLens could display was still too small for some participants. This issue was particularly problematic for the two participants with central vision loss, one of whom was completely unable to use the system to read (P5) and one of whom was frustrated by how little of his available vision could be used (P1). In contrast, another participant with tunnel vision found the display to be perfectly sized. The contrast of the HoloLens display also continued to be problematic for some participants, especially for P5 who was unable to see anything on the screen without blocking out all external light. These highly variable results re-emphasize the need for customizability.

Summary. Our second prototype system improved on several aspects of the first, with a better camera that could capture a greater amount of text, more polished and robust virtual display options, and efficient controls that allowed users to more easily customize the AR display to fit their needs. Participants were largely positive about our updated design, appreciating the options for customization and noting tradeoffs between the three AR display designs as well as advantages compared to

existing technology. The design sessions also helped to identify important features and design elements for future AR magnification aids.

5.3 Discussion

We reflect on the implications of our findings, focusing on ways to support a range of users with different visual impairments and a range of situations.

5.3.1 Overall Experience with 3D Augmented Reality

Our design sessions explored a novel AR magnification approach. The ability to display content in 3D space enables new interactions that are not possible with handheld devices or head-mounted 2D displays. For example, participants liked that they could use stationary 3D designs to create and position an arbitrarily large virtual display and then read a full document by turning their head, rather than scanning line by line as with other portable reading aids. Participants also liked how the design allowed them to easily multitask, for example, by turning away from the virtual display to speak with someone, then turning back to continue reading.

However, some participants commented that our approach required more effort to use than other reading aids. These participants preferred the simplicity of designs that fixed the display in 2D in front of their vision (*e.g.*, as in [235]), especially when they are trying to concentrate on the content of what they are reading. Further refinements to our designs and additional time for the participants to practice using the system may have improved their impressions of the system, but it is also possible that

more traditional reading aids or simple 2D image enhancements may work better for some situations or users.

5.3.2 Reflections on Head-mounted AR vs. Handheld Tools

AR magnification has potential benefits compared to other magnification approaches, but also limitations that must be addressed to provide a compelling alternative to existing aids.

Usability. Smartphone magnifiers are portable and readily available but have limited screen size. Users can hold the phone close to their face to compensate, but that may be uncomfortable for extended periods. Current HMDs do not yet address these issues, but we expect that future iterations will be lighter-weight and provide a perceptually larger display. These physical advances should allow users to read more naturally than with a handheld magnifier.

Flexibility. Our approach separates the camera from the display, allowing users to find a comfortable reading position regardless of the location of the physical world object, and supports customization so that users can adapt the display to their visual needs or context.

Privacy and Discreet Use. Handheld magnifiers and smartphone apps offer portability but may require the user to hold the device close to their face to read, preventing discreet use. While current HMDs attract attention for other reasons (unusual, bulky), we expect that future designs will be smaller and less noticeable, and

that users will be able to use the magnification aid more privately and discreetly than with a handheld screen—a feature that one participant found particularly appealing.

Ergonomics. Physical strain and fatigue are potential problems for many portable magnification aids [68]. This was also a limitation of our prototypes, but future AR designs could use a smaller wearable camera that can be aimed separately from the display for maximum flexibility and comfort. Participants also noted ergonomic problems with the HoloLens, including weight and eyestrain. These issues are also present to some extent with other head-worn vision enhancement systems. Future HMDs will need to be smaller and more ergonomic with screens designed to support a wide range of vision levels.

5.3.3 Recommended Design and Future Work

Based on the design sessions, we propose design recommendations and key features for assistive AR devices.

HMD. Participants liked our use of a transparent display that did not block out external vision, unlike most existing HMD systems (*e.g.*, [235,238]). Therefore, an ideal system should use an optical see-through HMD, but in a more lightweight form factor than the HoloLens, with a larger field of view to better support users with limited central vision. However, if contrast cannot be sufficiently improved in future optical see-through HMD devices, a video display that blocks out external light may be a better choice for some low vision users (*e.g.*, P5). Future work should explore alternative display options and evaluate their suitability for different users and contexts.

Camera. Participants valued flexibility, comfort, a wide field of view, and hands-free use, suggesting the need for a wearable camera that can be aimed separately from the display. The finger-worn and handheld smartphone cameras that we tested did not meet these criteria, but neither do the head-worn cameras used in most existing commercial HMD systems (e.g., [238,240]). A head-worn camera should allow for movement and optical zoom independent of the headset so that target content can be captured without requiring users to precisely position their head. Separate motion of the camera and head is also crucial for allowing users to move their head to scan virtual content in 3D, an interaction which participants found intuitive and useful. This design would likely require the ability to detect the content a user is pointing toward so that it can be magnified (*e.g.,* similar to the interaction used by OrCam [159]). Future work should explore these camera options in more depth.

Virtual display. AR magnification systems should include multiple display options to support different users and situations. We encountered tradeoffs between designs, such as the ease of use and attention required, ability to multitask, usefulness for different situations (*e.g.*, reading a long document *vs.* products in a store), and technical complexity and robustness. The ability to anchor virtual content in 3D space in the physical world can support a more natural and flexible reading experience compared to existing 2D vision enhancement systems, but it is also more complicated to implement and may have a steeper learning curve for users. Future work should investigate new AR designs, such as the ability to place multiple displays with different

targets or magnification levels, and an option to automatically display enhanced content directly over the text (*e.g.,* on a page or sign).

User input. Feedback from our study suggests that future systems should likely not use a smartphone camera because of the physical coordination and strain it required, but AR systems could still include intuitive and familiar touchscreen controls. Display settings could be adjusted using a smartphone or a smartwatch alongside the headset, or via touch controls on the headset itself (*e.g.,* the touch slider on Google Glass [56]). Participants also requested voice controls for some options, as well as physical buttons for key settings (*e.g.,* toggling the display, adjusting brightness and magnification). Future work should evaluate the efficiency and usability of these options.

New features. Although we did not investigate them in this work, future systems should also include features to help users read more easily in different situations. For example, optical or digital image stabilization would ensure smooth video, and optical character recognition (OCR) could help enhance text readability (*e.g.,* by changing fonts, increasing the resolution, or removing other visual elements such as images). OCR would also enable text-to-speech and other audio features (*e.g.*, as provided by OrCam [159]) alongside visual enhancements, as requested by some participants. Existing systems include some of these features already, but future work should investigate the usability of AR with hybrid visual and audio feedback. And beyond the ability to magnify and read nearby printed text materials, participants also mentioned several other desirable applications, such as reading signs, recognizing

faces, and attending sports events; future work should also explore how best to use AR to support these applications.

5.3.4 Limitations

The HoloLens has a narrow and centrally located field of view (estimated at 30° × 17.5°), which did not work well for some users. The translucent "holograms" that the HoloLens displays are also low-contrast, and colors are distorted by the screen material. Two participants were unable to use the device to read due to these issues, and most of the others mentioned them as limitations as well. We did not evaluate alternative headsets, although we anticipate that future versions of the HoloLens or similar technology will be able to address these issues. Future work should consult vision experts to better assess design requirements and usability for specific visual impairments. While our design sessions were informative and helped identify important design features for AR magnification aids, our study was not controlled, included a relatively small number of participants (7 total), and did not quantitatively evaluate usability or reading speed and comprehension. Future work should investigate possible camera positions and virtual display designs in more depth, and also directly compare AR magnification aids against existing technology.

5.4 Summary

This chapter explored novel applications of AR to assist low vision users, applying recent technology that can anchor 3D virtual content in the physical world. We explored the AR magnification design space and presented two prototype systems that

we evaluated and refined through iterative design sessions with low vision participants. Participants liked the general concept of a head-worn magnification aid for its improved portability, privacy, and ready availability compared to other magnification aids they had used. Participants also identified advantages to our 3D AR approach compared to handheld magnification tools, including a more natural reading experience and the ability to more easily multitask, but also some disadvantages such as a steeper learning curve and limitations of the particular hardware we used. Through our open-ended design and evaluation sessions, we identified several common themes that should inform the design of future AR vision enhancement aids for low vision users.

Chapter 6: Localization of Skin Features on the Hand and Wrist

We have demonstrated the feasibility of using finger-mounted sensing and feedback to enable blind users to access printed text materials, but our approach could also potentially support many additional applications. This chapter explores the preliminary algorithmic foundations necessary to support one such application: *on-body interaction*, an emerging paradigm in HCI where users tap or gesture on their own body to control a mobile device and access digital information (*e.g.*, [40,69,70,117,139,151, 153,209,223]). One advantage of this type of input is that it is always available, allowing the user to, for example, quickly tap or swipe on their palm to answer a phone call or listen to new emails (Figure 6.1a). On-body interaction is also useful when visual attention is limited because the skin's tactile perception allows for more accurate input than is possible with a touchscreen [64,154].

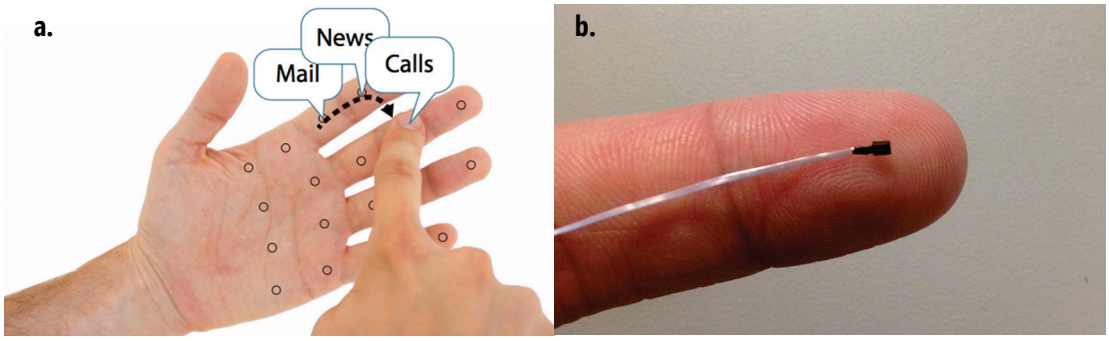


Figure 6.1: (a) Conceptual visualization of on-hand input to control a mobile phone, as in [64]. (b) Cameras developed for minimally invasive surgeries are small enough to mount on the finger. Shown: AWAIBA NanEye ($1\times1mm^2$, 250×250px resolution) used in [199,228].

This chapter contains work published in the proceedings of the 23rd International Conference on Pattern Recognition (ICPR 2016) [202].

Sensing these on-body taps and gestures, however, is a challenging problem. Researchers have investigated a variety of wearable cameras (*e.g.*, [65,70]) and other sensors (*e.g.*, bio-acoustics [69], ultrasonic rangefinders [117]). While promising, these approaches are limited by the placement and range of the sensor [65,151], suffer from occlusion [65] or precision [69] problems, or cover the user's skin [220], reducing tactile sensitivity. Instead, we envision using close-up images from a small *finger-mounted camera* (*e.g.*, [199,228]) to sense and localize user input (Figure 6.1b). By instrumenting the gesturing finger with a camera, our approach extends the user's interaction space to anything within reach and can support precise location-based input.

Previous work in skin classification has largely been in the context of biometrics—that is, determining the *uniqueness* of a user's skin patterns for identification purposes (*e.g.,* [30,39,46,85,136,141,225]). In this chapter, rather than identifying *who* an image represents, we seek to identify *where* an image is located on a single user's body. More specifically, we investigate to what extent are surface image patches of the hand and wrist localizable?

Localizing small (~1–2 cm) image patches within the larger skin surface is similar to partial finger and palm print recognition in forensic applications; however, high-resolution, high-contrast images of ridge impressions are typically needed to reliably extract distinctive point and line features. In contrast, cameras small enough to be mounted on the finger (Figure 6.1b) are low resolution and low contrast, making it difficult to detect minute ridge features. Several recent biometric systems recognize finger and palm prints using lower-quality images [31,40,43,45,94,141,159,163,218,

227]. Unfortunately, these approaches are frequently designed to align and process the finger or palm image as a whole and cannot reliably recognize a small portion of the print. To our knowledge, no work has attempted to recognize or localize a small skin patch from live camera images, which we do here.

To ultimately support on-body localization using a finger-mounted camera, we investigate the classifiability of 17 locations on the front and back of the palm, fingers, wrist, nails, and knuckles. We introduce a hierarchical texture classification approach to first estimate the touch location on the body given close-up images of the skin surface and then refine the location estimate using keypoint matching and geometric verification. To evaluate our approach, we collected a skin-surface image dataset consisting of 30 individuals and the 17 hand and wrist locations (10,198 total images).

When testing and training on an individual's own skin data (within-person experiments), our results show that skin patches are classifiable by location under controlled conditions with 96.6% recall and 96.4% precision, suggesting that finger-mounted cameras may be feasible for sensing on-body interactions.

In summary, the contributions of this chapter include: (i) a robust algorithmic pipeline for recognizing several different locations on the hand from small patches of skin; (ii) classification results for a dataset consisting of 30 individuals, achieving accuracy above 96% on average for within-person experiments; and (iii) analysis of hand distinctiveness and similarities among users, which may impact accuracy and scalability (*e.g.,* between-person training feasibility).

6.1 Touch Localization Pipeline

Robust localization of close-up skin images from a finger-mounted camera is challenging due to the limited field of view (~1–2 cm) and relatively low contrast of the ridges and other skin surface features. To estimate the user's touch location from close-up images, we developed a hierarchical classifier with four stages: (i) preprocessing, (ii) coarse-grained classification, (iii) fine-grained classification, (iv) geometric verification and refinement. The coarse-grained stage classifies an input image into one of five regions: *palm*, *fingers*, *nail*, *knuckle*, and *other* (wrist and back of hand). The fine-grained stage further classifies the image into a discrete location within that region (17 locations in all; see Figures 6.2 and 6.5). These locations were selected because previous work has shown that users can reliably locate them with high accuracy even without sight [154]. While our four-stage pipeline integrates multiple known approaches in fingerprint and palmprint enhancement, texture classification, and 2D keypoint matching, our primary innovation is in their novel combination and application towards *localization* rather than identification.

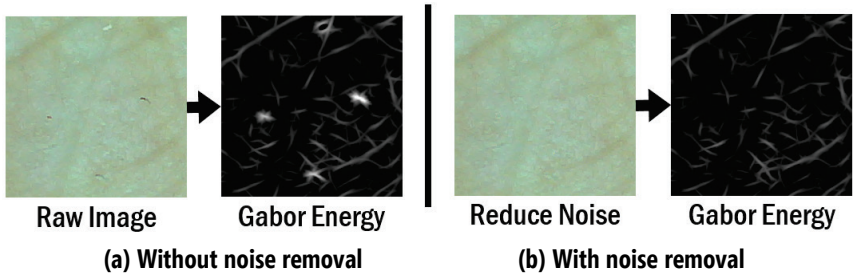


Figure 6.2: Stage 1 preprocessing first removes dirt and other noise before emphasizing ridge features using the energy of a set of Gabor filters with different orientations. Shown: an example image from the left side of the palm, scaled and cropped to demonstrate the effect that surface artifacts can have on the Gabor energy image.

Stage 1: Preprocessing. Images are first preprocessed to remove noise and emphasize ridge features. We apply an efficient median filter [88] to reduce the effect of dirt and other camera noise while preserving the edge information necessary for processing finger and palm prints (Figure 6.2).

To emphasize the ridgelines, we adapt a technique from Huang *et al.* [80]. However, while they use a modified version of the finite radon transform to emphasize the principal lines and creases of the palm, these features are not as prominent in our images due to the narrow field of view. We instead use Gabor filters. We compute the Gabor energy image defined as the maximum response at each pixel from a set of Gabor filters with different orientations. Specifically, the energy at pixel location (*x, y*) is:

$$E_{x,y} = \left| \max_{\theta} \left[G_\theta * \left(\bar{I}_{x,y} - I_{x,y} \right) \right] \right| \quad (1)$$

where $I_{x,y}$ is the gray-scale pixel value at (*x, y*) and $\bar{I}_{x,y}$ is the local mean in a window around that location (estimated using a Gaussian smoothing function), G_θ is a discrete Gabor filter with orientation θ, and * is the convolution operator. In our experiments, we use 18 uniformly distributed orientations, with a fixed scale and bandwidth that were chosen empirically based upon the average ridge frequency in our preliminary experiments with a separate set of pilot data. Example energy images are shown in Figures 6.2, 6.3, and 6.4.

Stage 2: Coarse-Grained Classification. After preprocessing, we obtain a rough classification of the image's location using the visual texture, which we represent using LBP histograms. We chose LBP because of its computational efficiency and

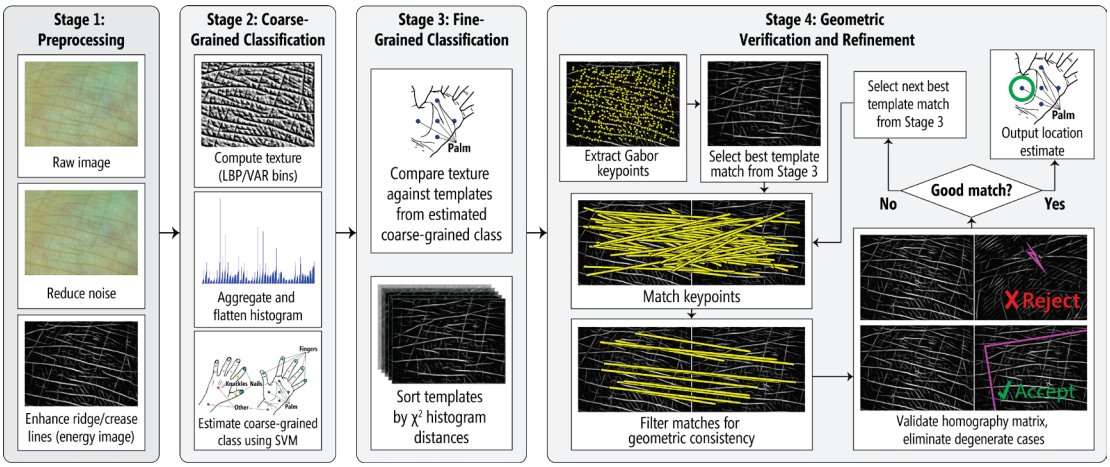


Figure 6.3. The four stages of our localization algorithm, as applied to an example image from the left side of the palm. First, the image is preprocessed to remove surface artifacts and camera noise before calculating the Gabor energy to emphasize ridge and crease lines. Second, the image is classified into one of five coarse-grained locations (in this case, the palm) using a 2D texture histogram of LBP and pixel variances. Third, the image's texture is compared against the templates from the predicted coarse-grained class, which are sorted by their χ^2 histogram distances to prioritize matching for the next stage. Finally, the image is compared geometrically against images from the predicted coarse-grained class, using a set of custom Gabor keypoints and descriptors. The image is compared against individual templates starting with the most likely match (as predicted in Stage 3), proceeding in order until a template with sufficient geometrically consistent keypoint matches is found. If a geometrically consistent match is found, then the fine-grained location can be estimated with a high degree of certainty (in this case, the left side of the palm); otherwise, the algorithm falls back upon the closest texture match from Stage 3.

natural invariance to illumination variations. To improve accuracy and achieve rotation invariance, we use only the uniform patterns alongside the variance of the neighboring values as suggested in [157]. Our implementation uses a 2D histogram with 14 uniform pattern bins and 12 variance bins ($LBP^{riu2}_{12,2}$ and $VAR_{12,2}$, as defined in [157]), computed at 3 scales. These parameters were selected because they provided a balance between classification accuracy and computational efficiency on our pilot data. The histograms for each scale are flattened and concatenated together to produce a 672-element feature vector, which is then normalized. To classify the LBP histograms into coarse-grained

body regions, we train a support vector machine (SVM)—commonly used in texture classification (*e.g.*, [42,100,228]).

Stage 3: Fine-Grained Classification. We compare the LBP histogram using a template matching approach against *only* the training templates from the coarse-grained region identified in Stage 2. This hierarchical approach reduces the number of possible match locations and enables us to prioritize different features for each region individually (*e.g.*, for the palm we can automatically weight the palmprint texture features that best discriminate the five different palm locations). For template comparisons, we use the χ^2 distance metric, which is known to perform well with LBP histograms (*e.g.,* [4]). Stage 3 produces a sorted list of templates, with the lowest distance representing the most likely match.

Stage 4: Geometric Verification and Refinement. Stage 4 ensures the validity of the texture match and refines the precise touch location using a set of keypoint matching and geometric verification steps. We investigated SIFT keypoints

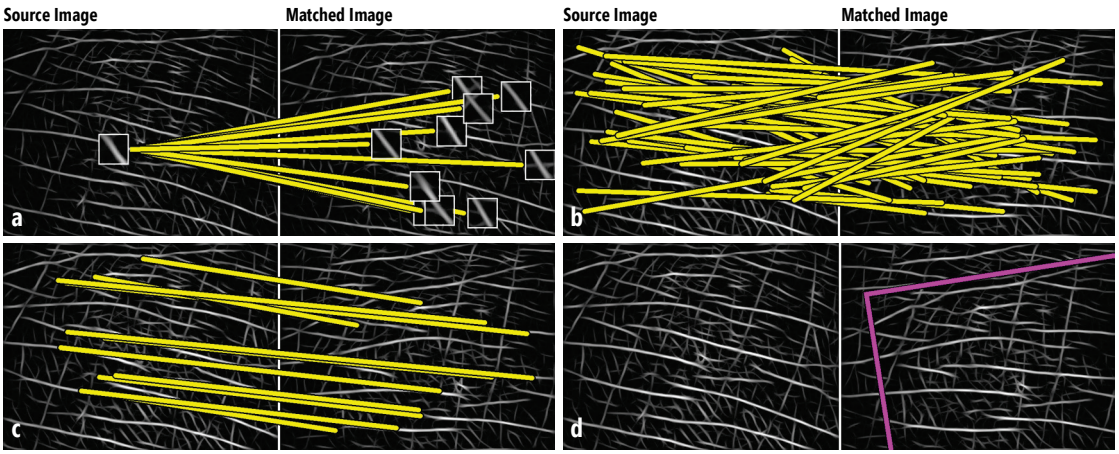


Figure 6.4: Keypoints in the Gabor energy images frequently appear visually similar (a), leading to a high percentage of mismatches (b). We filter outliers using a series of verification steps to ensure geometric consistency (c and d).

[141,162,225], Harris corners [162], and fingerprint minutiae [46,85,106], but found them too unreliable in preliminary tests. Instead, we use keypoints with a high Gabor filter response at two or more orientations, which tend to lie at the intersections of ridgelines or creases. The Gabor energy values in the 16×16px neighborhood surrounding the keypoint serve as a reliable descriptor. To achieve rotation invariance, we generate multiple descriptors at each keypoint location, rotating the neighborhood for each using the orientation of the filters with locally maximum response strength. We keep a list of keypoints for each training image.

These image patches, however, are frequently visually similar (*e.g.*, Figure 6.4a), leading to a high percentage of mismatches between the keypoints (Figure 6.4b). We address this issue using a series of geometric verification steps. First, we filter the matches for orientation consistency, eliminating matches that do not agree with the majority vote for the relative rotation between images (*i.e.*, any more than 20° from the average rotation across all matches). Second, we compute a homography matrix using random sample consensus (RANSAC), identifying inliers and ensuring that there are sufficient geometrically consistent feature matches (*i.e.,* more than the minimum necessary to define a homography; in our experiments, we required 16 consistent matches). Although the palm and fingers are not rigid planar surfaces, in the close-up images we gathered they appear nearly so; we compensate for irregularities by allowing a greater than usual inlier distance of 10 pixels. Third, we verify that the homography matrix is well behaved using the following constraints, which ensure that the match preserves orientation and does not have extreme variations in scale or perspective:

$$
\begin{aligned}
&1.\quad H_{11}H_{22} - H_{21}H_{12} > \tfrac{1}{2} \\
&2.\quad \tfrac{1}{2} < \sqrt{{H_{11}}^2 + {H_{21}}^2} < 2 \\
&3.\quad \tfrac{1}{2} < \sqrt{{H_{12}}^2 + {H_{22}}^2} < 2 \\
&4.\quad \sqrt{{H_{31}}^2 + {H_{32}}^2} < \tfrac{1}{1000}
\end{aligned}
\tag{2}
$$

These constraints were selected empirically to eliminate most degenerate cases that could lead to false-positive matches. Fourth and finally, to avoid further degenerate cases, we ensure that the inlier features are not collinear and that they have sufficient spread. We do this by calculating the standard deviation along the two principal directions computed using principal component analysis; if $\sigma_1 < 25$ or $\sigma_1/\sigma_2 > 4$, we declare the match invalid (these numbers were also selected empirically and validated on a separate set of pilot data). If a template match is declared invalid, we proceed to the next best texture match, stopping once we find one that passes all conditions. If a valid match is not found, then we fall back upon the best Stage 3 texture match.

The output of our hierarchical algorithm is an estimated classification of a query image into one of 17 locations, along with a confidence score based upon the texture similarity and the number of inliers for the best template match. From the computed homography matrix, we also obtain a more precise location estimate relative to the matched templates, potentially enabling finer localization for future explorations.

6.2 Data Collection and Dataset

To evaluate our approach, we created an image dataset collected from 30 volunteers (7 male, 23 female) recruited via campus email lists. The participants were on average

Participant Demographics

Gender	23 female, 7 male	
Age	Mean = 30.6, SD = 11.5, Min = 18, Max = 59	
Race	Black, Afro-Caribbean, or Afro-American	6
	East Asian or Asian-American	5
	Latino or Hispanic American	1
	Non-Hispanic White or Euro-American	14
	South Asian or Indian American	2
	Other or Multiple	2
Palm Size	Mean = 98.3 mm, SD = 10.3 mm, Min = 79.7 mm, Max = 129.5 mm	

Table 6.1: Our dataset captures variations in gender, age, race, and palm size. Palm size was measured diagonally from the base of the thumb to base of the smallest finger while the fingers were spread and fully extended.

30.6 years old (*SD*=11.5, *range*=18–59), and represented a variety of skin tones and palm sizes (Table 6.1). For each participant, we collected close-up images of 17 locations (Figure 6.5) using a small 0.3MP micro-lens camera in the shape of a pen.

The micro-lens camera is self-illuminated with a manually adjustable focal length, enabling us to capture clean 640×640px images of the hand from as close as 1cm. We controlled for distance and perspective using two 3D-printed camera attachments that place the camera approximately 2.5cm from the surface of the hand, at either a 90° or 45° angle (Figure 6.5b). Compared to a finger-mounted camera, this form factor enabled us to more easily control for variables such as distance, perspective, focus, and lighting, while still capturing images that are representative of our target domain. Ultimately, we expect to use a smaller camera similar to Figure 6.1b.

Participants used the camera to point to 17 locations on the hand and palm, with 10 trials for each location and two perspectives (45° and 90°) resulting in 340 images per person. Rather than point 10 times in a row to the same location, the order of trials was randomized to provide variation in translation, rotation, and pressure (which

impacts scale and focus). In total, we have 10,198 close-up micro-lens images across the 30 participants (one participant skipped two trials). While we would like to release this dataset publicly, we cannot do so without risking the privacy of our participants.

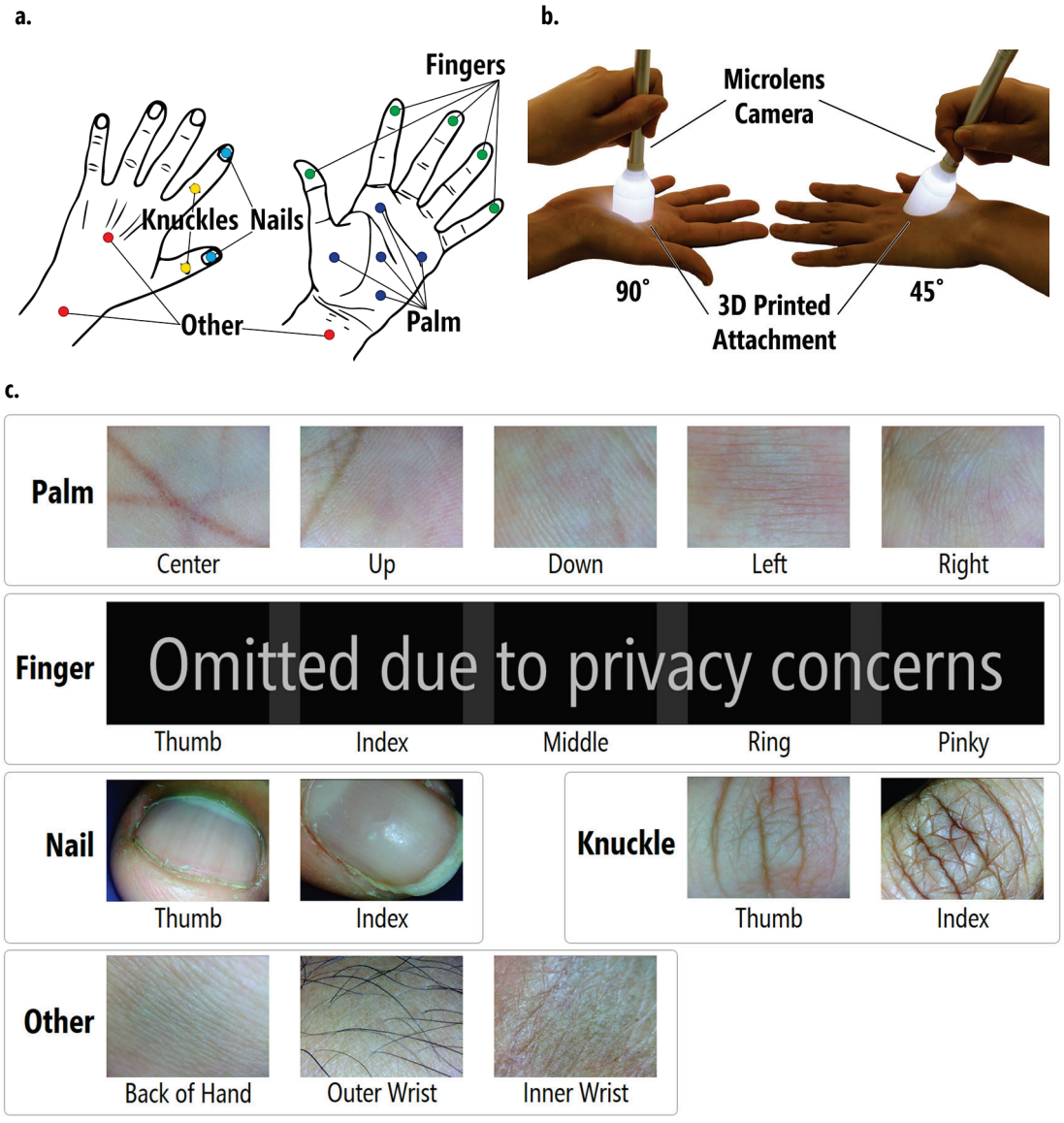


Figure 6.5: Data collection setup: (a) 17 close-up image locations on the left hand in 5 coarse-grained regions–coded with different colors; (b) the pen-based camera and physical constraints (one angled at 45° and one at 90°) used for close-up image capture. (c) representative images from our dataset for each of the 17 locations, selected across 12 participants.

6.3 Experiments and Results

We first describe results related to coarse- and fine-grained hand classification performance before presenting secondary analyses related to the effect of training sample size on performance and between-person classification. Our analyses report standard measures including precision, recall, and F_1 scores. These metrics are more informative than accuracy due to the uneven number of training examples per class our hierarchy defines.

6.3.1 Within-Person Classification

To evaluate the overall location-level classifiability of the hand, we conducted a within-person experiment. We used an *n*-fold, leave-one-out cross-validation approach. Our results are the average across all 20 folds for each of the 30 participants. We first present aggregate results before examining performance by location and by participant.

At the coarse-grained level (Stage 2), the average precision is 99.1% (*SD*=0.9%) and average recall is 99.2% (*SD*=0.8%). At the initial fine-grained level (Stage 3), the average precision is 88.2% (*SD*=4.4%) and recall is 88.0% (*SD*=4.5%). After performing geometric validation and refinement (Stage 4), fine-grained classification increases to 96.6% precision (*SD*=2.2%) and 96.4% recall (*SD*=2.3%). The high precision and recall values demonstrate the feasibility of using close-up images to classify locations on the hand and wrist. Stage 2 precision and recall are very high (above 99%), which is important because errors in estimating the coarse-grained region will propagate to the next stage (a limitation of our hierarchical approach).

Stage 2: Coarse-grained Classification Confusion Matrix

	Palm	*Finger*	*Nail*	*Knuckle*	*Other*
Palm	**99.0%**	0.5%			0.5%
Finger	0.6%	**99.3%**	0.1%		
Nail	0.2%	0.1%	**99.7%**	0.1%	
Knuckle			0.2%	**99.1%**	0.7%
Other	0.6%		0.1%	0.5%	**98.8%**

Table 6.2: Classification percentages for classes at the coarse-grained level. Each cell indicates the percentage of images assigned to a predicted class (column) for each actual class (row).

Across all stages, we observed classification errors that were caused primarily by similarities between the locations' visual textures, poor image quality, and insufficient overlap between the training and testing images, although the high accuracies meant that there was not enough data for statistical analysis of the errors.

To examine the impact of different hand/wrist locations on performance, we created confusion matrices for Stage 2 (coarse-grained) and Stage 4 (fine-grained) classifications. See Tables 6.2 and 6.3 respectively. The locations with the lowest F_1 score were those on the back of hand (M=92.3%; SD=10.1%) and wrist (M=91.8%;

Stage 4: Fine-grained Classification Confusion Matrix

	Palm					*Fingers*					*Nails*		*Knuckles*		*Other*		
	C	*U*	*D*	*L*	*R*	*1st*	*2nd*	*3rd*	*4th*	*5th*	*1st*	*2nd*	*1st*	*2nd*	*BH*	*OW*	*IW*
Palm Center (C)	**98.3%**					0.2%		0.2%		0.2%					0.2%	0.5%	0.5%
Palm Up (U)	0.2%	**98.5%**		0.2%	0.2%		0.2%			0.3%						0.2%	0.3%
Palm Down (D)	0.3%	1.2%	**95.7%**	0.2%	1.7%	0.3%	0.2%						0.2%			0.2%	0.2%
Palm Left (L)	0.3%	0.3%	0.2%	**98.7%**	0.3%	0.3%											0.7%
Palm Right (R)	0.7%	0.5%	0.3%	0.5%	**97.5%**	0.2%	0.2%			0.2%							
1st Finger		0.5%	0.2%	0.2%	0.7%	**96.3%**	0.3%	0.5%	0.5%	0.7%					0.2%		
2nd Finger		0.3%			0.2%	0.3%	**95.8%**	1.7%	0.5%	1.2%							
3rd Finger			0.3%			0.2%	1.3%	**95.4%**	2.2%	0.7%							
4th Finger			0.2%				0.3%	1.8%	**95.3%**	2.3%							
5th Finger		0.2%		0.2%			0.3%	0.5%	1.5%	**97.0%**	0.3%						
1st Nail				0.2%							**98.2%**	1.7%					
2nd Nail			02%						0.2%		0.5%	**99.0%**		0.2%			
1st Knuckle											0.2%		**97.3%**	1.2%	0.2%	0.2%	1.0%
2nd Knuckle												0.2%	0.8%	**98.8%**		0.2%	
Back of Hand (BH)				0.2%									0.5%	0.2%	**92.2%**	4.7%	2.3%
Outer Wrist (OW)	0.2%												0.2%		6.0%	**90.2%**	3.5%
Inner Wrist (IW)	0.7%	0.7%	0.2%	0.2%	0.3%						0.2%		0.7%	0.2%	0.8%	0.5%	**96.2%**

Table 6.3: Classification percentages for classes at the fine-grained level (Stage 4 output), averaged across 20 trials and 30 participants. Each cell indicates the percentage of images assigned to a predicted class (column) for each actual class (row).

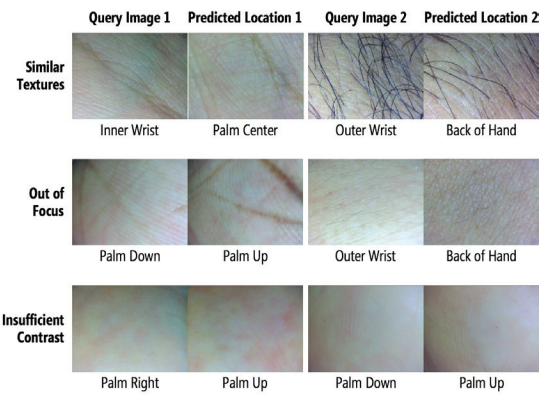


Figure 6.6: Classification errors were caused primarily by similarities between the locations' visual textures and poor image quality. Each set of images shows, in order, two examples (from different participants) of an incorrectly classified test image along with a training image from the predicted location.

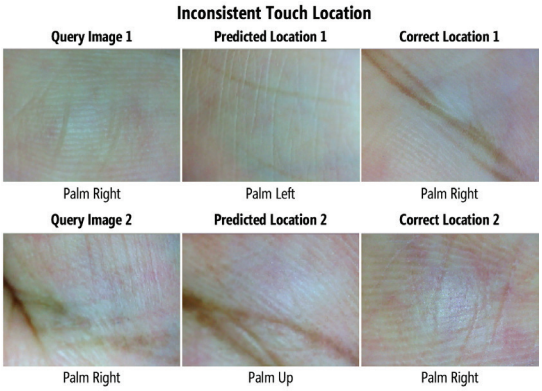


Figure 6.7: Classification errors for several participants were also caused by inconsistent touch locations. Shown are two examples of query, predicted, and correct locations (from two different participants) where the touched locations were far enough apart to appear as entirely unrelated images.

SD=8.4%), which appear visually similar (Figure 6.6). This was true to a lesser extent across all coarse-grained regions, with the textures of different locations within each region appearing similar. While Stage 4 geometric validation reduced misclassifications, it was not always successful. For example, in some cases, an image for a participant did not sufficiently overlap any other image in the dataset, preventing geometric keypoint matching (Figure 6.7). In these cases, the algorithm fell back to the best Stage 3 texture match.

To examine how performance varies across individuals, Figure 6.8a shows F_1 scores broken down by participant. F_1 scores ranged from 95.9% to 100.0% at the coarse-grained level (Stage 2) and 86.5% to 99.7% at the fine-grained level (Stage 4). Participant 29 performed the worst, with a Stage 4 F_1 score of 86.5%—4.4 standard deviations below the mean. Based on a qualitative examination, we found decreased skin contrast with fewer distinctive finger and palm features, as well as significant

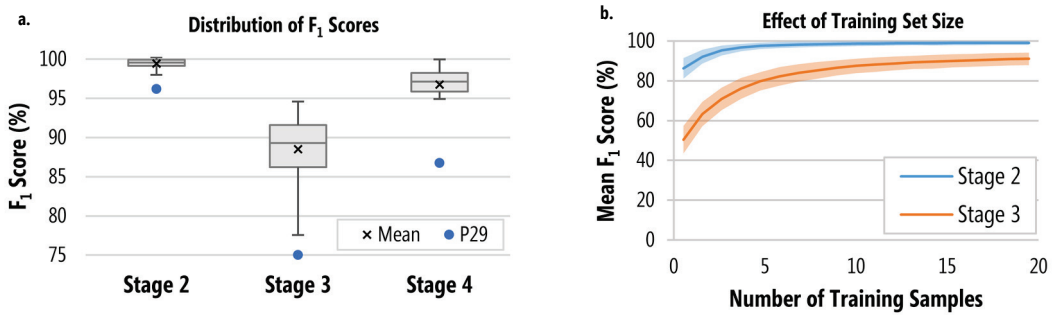


Figure 6.8: (a) Distribution of F_1 scores by participant, with outlier P29 marked by the blue dot; (b) Effect of the number of training examples on mean texture classification F1 score at coarse-grained (Stage 2, blue) and fine-grained (Stage 3, orange) levels.

variations in translation, rotation, and image focus for each location. In comparison, the top performing participants had high contrast skin textures, more consistent pressure (resulting in fewer variations in lighting and focus), and greater consistency in returning to the same touch location each trial.

6.3.2 Effect of Training Set Size on Performance

To explore performance as a function of training set size, we tested our algorithms again using *n*-fold cross-validation but this time varying the number of training samples from $m = 1$ to 19. Specifically, we randomly selected from the 20 images per class available for each participant, with one image set aside for testing. Figure 6.8b shows the average texture classification accuracy at the coarse-grained (Stage 2) and fine-grained (Stage 3) levels when increasing the number of training examples. To reduce the effect of selecting the images randomly and obtain a more representative estimate, we averaged the results of 10 randomized trials. Each point represents the average F_1 score across all participants, locations, and trials when trained using m examples. Accuracy begins to level off above five training images per location, especially at the coarse-grained level (which approaches 100% accuracy). However, performance at

both levels steadily improves as the number of training images is increased. We did not evaluate Stage 4 for this experiment as its performance depends largely upon the amount of spatial overlap between the training and testing images rather than the number of training samples.

6.3.3 Between-Person Classification

To potentially bootstrap the training set and to identify similarities across individuals, we conducted a secondary classification experiment in which the training set and testing set consisted of images from different participants (*i.e.,* between-person experiments). More specifically, we employed *n*-fold cross-validation, where each fold trained on data from 29 participants and tested on the remaining participant. We did not expect this approach to yield a high accuracy, especially at the fine-grained level since finger and palm prints can vary significantly person to person (which is the basis of biometric identification). However, we hoped to discover textural similarities across participants that could be used to boost future classifiers to either improve accuracy or reduce the amount of per-user training.

As expected, the between-person classification results are lower than the within-person results. At the coarse-grained level, our classification algorithms achieve

Between-person Coarse-grained Classification Confusion Matrix

	Palm	*Finger*	*Nail*	*Knuckle*	*Other*
Palm	**55.2%**	16.8%	7.8%	4.0%	38.8%
Finger	8.1%	**85.5%**	10.4%	2.1%	2.3%
Nail	0.2%	3.4%	**85.3%**	4.4%	0.9%
Knuckle	1.2%	0.2%	1.3%	**67.8%**	18.2%
Other	12.4%	4.1%	0.1%	18.2%	**60.3%**

Table 6.4: Between-person classification percentages for classes at the coarse-grained level. Each cell indicates the percentage of images assigned to a predicted class (column) for each actual class (row).

an average precision of 72.6% (*SD*=12.9%) and recall of 70.8% (*SD*=12.3%). Still, these results are considerably higher than chance for five classes (20%) or majority-vote for the palm class (29.4%). See Table 6.4 for a confusion matrix. Average precision at the fine-grained level is 27.1% (*SD*=7.5%) and recall of 26.1% (*SD*=5.8%), which are also well above chance for 17 classes (5.9%). Although these accuracies are clearly too low to support a reliable user interface without an individual training procedure, they may provide enough information to allow for bootstrapping.

6.4 Discussion

Our controlled experiments explored the distinguishability of small image patches on the surface of the hand and wrist for localization purposes. In our within-person experiments we were able to achieve an average F_1 score above 99% at the coarse-grained level (Stage 2) and above 96% at the fine-grained level (Stage 4), which suggests that skin-surface image patches can be classified and localized on the body with high levels of accuracy. While an end-to-end deep learning approach may be more elegant, our more heuristic approach requires substantially less training data, and our performance results suggest that an on-body input system applying our algorithms is feasible. Here, we reflect on the implications of our findings as well as challenges for implementing a real-time system.

6.4.1 Expanding On-Body Input

While we only evaluated locations on the hand and wrist, our finger-mounted approach should support a range of input locations within the user's reach, including on-body

and off-body surfaces (*e.g.*, tabletops). This is in contrast to most previous on-body input approaches that are more limited by their fixed sensor placements and range. Although recognition accuracy may drop as the number of locations increases (*e.g.*, thigh, forearm), we expect to boost performance through improvements to our hierarchical approach. Performance was particularly high at the first level of the hierarchy, with an F_1 score above 99%. Thus, for each region we could apply different preprocessing and matching approaches at the second level that are tuned specifically to distinguish the fine-grained locations within that region. For example, we could extract knuckle-specific features (*e.g.*, [30]) to distinguish knuckle locations, which may require completely different algorithms than the palm locations. Similarly, it will be important to explore the feasibility of extending the localization hierarchy further, for regions that can support an even finer level of granularity beyond the locations studied (*e.g.*, palm, fingers); such granularity could enable highly precise on-body interactions (*e.g.*, sliding your finger along your palm to trace a map route).

6.4.2 Training a Camera-Based On-Body Localization System

The procedure for training a new user may impact both algorithmic performance and user perceptions toward the system. As shown in Figure 7b, classification performance improves with the number of training examples, but begins to level off after five examples per class. However, it may be possible to boost accuracy while simultaneously reducing the number of training examples that are required of a new user. The images in our dataset relied on natural variations that were introduced through randomization during data collection. To potentially improve performance, the training

interface could prompt the user to vary rotations, poses, and perspectives—similar to Apple's iPhone training procedure for their fingerprint sensor. In addition, as our preliminary experiments indicate, it may be possible to bootstrap the system using between-person data and reduce the amount of training required for a new user. This approach would work especially well in our first stage of classification to recognize surface classes that appear similar across many users (*e.g.*, skin, knuckles, clothing).

6.4.3 Limitations and Future Work

Our experiments were conducted under controlled conditions, but a real-time system would likely need to deal with greater variations in image quality. Although we randomized trial order to introduce natural variation in translation, rotation, and pressure, we carefully controlled for other variables such as distance, lighting, and perspective. A finger-worn camera will likely constrain this complexity, potentially mitigating these concerns. For example, distance will remain relatively constant during touch-based interactions since the camera can be positioned at a fixed location on the finger and lighting can be controlled via a self-illuminated camera. While perspective may vary considerably, our results show that our algorithm functions well for both 90-degree and 45-degree perspectives. Further work is necessary to explore variations under less controlled conditions, including potential changes over time (*e.g.*, due to differences in humidity/dryness), as well as other variations in skin surface textures and features due to age, skin tone, and hand size. The above mitigating factors suggest that our approach should still be applicable.

Our work focused solely on RGB camera-based sensing using static images. Future research should explore other imaging and non-imaging sensors as well as combining video and multiple sensor streams (sensor fusion). For example, hyperspectral imaging would expose veins and other sub-dermal features that could be used for localization as well as improve the contrast of surface features across a wider range of skin tones (*e.g.*, [53]). Depth sensors could provide 3D geometry of the hand and ridges, potentially improving robustness to variations in perspective and allowing us to more reliably extract finger and palm print features to use for localization (*e.g.*, [116]). Finally, non-imaging sensors (*e.g.,* infrared reflectance [151] or inertial motion [86]) could provide complementary information to help resolve visual ambiguities and better integrate localization with gesture recognition.

6.5 Summary

This chapter introduces an algorithmic pipeline for recognizing low-resolution, close-up images of several different locations on the hand/wrist with an average F_1 score of 96.5% for within-person skin patch classification. While future work will need to address potential implementation challenges with a real-time system, our results suggest that a finger-mounted computer vision approach to support location-based on-body interaction should be feasible and that the system training process may be able to be bootstrapped using a dataset of hand images collected from multiple individuals.

Chapter 7: Realtime Recognition of Location-Specific On-Body Gestures

As discussed previously, on-body interaction offers several advantages over existing touchscreen devices. Taps, swipes, or other on-body gestures provide lightweight and always-available control (e.g., [63,70]) with an expanded input space compared to small-screen wearable devices like smartwatches (e.g., [109,118,150,152]). In addition, the proprioceptive and tactile cues afforded by on-body input can improve eyes-free interaction (e.g., [40,119,131,218]) and enable accurate input even without visual feedback compared to the smooth surface of a touchscreen [64,154]. These advantages are particularly compelling for users with visual impairments, who do not benefit from visual cues and who frequently possess a heightened sense of tactile acuity ([55,149]).

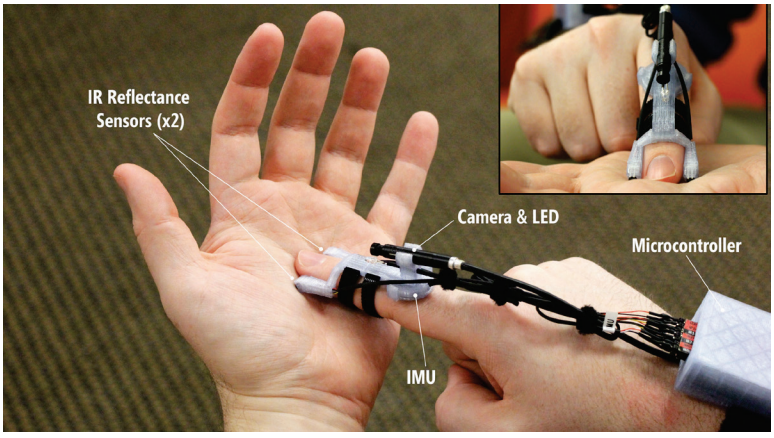


Figure 7.1: *TouchCam* combines a finger-worn camera with wearable motion trackers to support location-specific, on-body interaction for users with visual impairments. See supplementary video for a demonstration: https://youtu.be/VREiWI_38BQ.

This chapter contains work published in the Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT Dec 2017) [203].

Reliably sensing on-body input, however, is still an open challenge. Researchers have explored a variety of approaches such as cameras (e.g., [25,70,195]), infrared-reflectance sensors [97,109,150], and bio-acoustics [69,110]. While promising, this prior work has not been specifically designed for or tested with visually impaired users, who likely have different needs and preferences. For example, blind users may encounter difficulty accurately aiming a camera (or other directed sensor) [86,213] and also rely more on their sense of touch [55,149] making it especially important to avoid covering the fingertips. Furthermore, prior work does not support complex gestures at multiple body locations. For example, *Skinput* [69] detects touches at a range of locations but not more complex gestures. In contrast, *FingerPad* [27] and *PalmGesture* [218] sense shape gestures performed on the fingertip or palm but cannot easily be extended to other locations.

Our research explores an alternative approach using an updated finger wearable prototype to support location-specific, on-body interaction. We refer to this updated prototype as *TouchCam* (Figure 7.1) in this chapter to differentiate it from earlier work. The previous chapter demonstrated the feasibility of recognizing body locations from small skin surface images (1–2 cm) captured using a handheld camera; however, this work did not include sensor fusion, use a wearable form factor, or function in real-time. In addition, the previous prototype could only recognize locations (not gestures), and the images were collected under carefully controlled conditions. In this chapter, we build on that work and address these limitations. TouchCam combines data from infrared reflectance (IR) sensors, inertial measurement units (IMU), and a small camera

to classify body locations and gestures using supervised learning. Because TouchCam instruments the *gesturing* finger, on-body interaction is supported on a variety of locations within the user's reach while also mitigating camera framing issues. TouchCam also enables new location-specific, contextual gestures that are semantically meaningful (*e.g.,* tapping on the wrist to check the time or swiping on the thigh to control a fitness app). These features allow for flexible interface designs that can be customized based on the needs of the application or user. In this chapter, we explore four high-level research questions:

RQ1. How well can we recognize location-specific on-body gestures using finger- and wrist-mounted sensors?

RQ2. Which locations and gestures can be recognized most reliably using this sensing approach?

RQ3. What tradeoffs must be considered when designing and building a realtime interactive on-body input system?

RQ4. How accessible is this approach to users with visual impairments and what are their design preferences?

To address these questions, we evaluated two prototype iterations across two studies. In Study I, we demonstrate feasibility through a controlled data collection study with 24 sighted participants who performed touch-based gestures using the first iteration of our prototype (*TouchCam Offline*). In offline experiments using classifiers trained per-user, we achieve 98% accuracy in classifying coarse-grained locations (*e.g.*, palm, thigh), 84% in classifying fine-grained locations (*e.g.*, five locations on the

palm), and 96% in classifying location-specific gestures. Informed by these results, we built a second prototype with updated hardware and software algorithms to support *realtime* on-body localization and gesture recognition (*TouchCam Realtime*). In Study II, we investigate the usability and potential of the realtime system with 12 blind and visually impaired participants. Our findings validate realtime performance with our target population and highlight tradeoffs in accuracy and user preferences across different on-body inputs.

In summary, the primary contributions of this chapter include: (i) two iterations of TouchCam, a novel finger-worn camera system that uses machine learning to detect and recognize on-body location-specific gestures; (ii) a quantitative evaluation of our system's accuracy and robustness across a variety of gestures and body locations; (iii) qualitative observations about the usability and utility of our on-body input approach for users with visual impairments; and (iv) design reflections for on-body gestural interfaces in terms of what locations and gestures can be most reliably recognized across users. While our prototype design is preliminary—we expect that future iterations will be much smaller and self-contained—our explorations build a foundation for robust and flexible on-body interactions that support contextual gestures at multiple body locations via supervised learning. Our primary focus is supporting users with visual impairments; however, our approach could also benefit users with situational impairments (*e.g.*, while walking or conversing) or be applied as an input mechanism for virtual reality systems (*e.g.*, for accurate touch-based input in eyes-free situations).

All software code and hardware design files are open sourced and available here: https://github.com/lstearns86/touchcam.

7.1 TouchCam Offline: Initial Wearable Prototype

We describe our first prototype, TouchCam Offline, which we evaluate offline using data collected from a controlled study. Study I focuses on addressing RQ1 and RQ2: *how accurately can we recognize location-specific on-body gestures* and *which locations and gestures can be recognized most reliably?* Our results inform the development of a realtime prototype, which is evaluated in Study II (Section 7.3).

7.1.1 Prototype Hardware

The TouchCam Offline hardware consists of: (i) a finger-worn multi-sensor package that includes two infrared (IR) reflectance sensors, an inertial measurement unit (IMU), and a small camera with an adjustable LED for illumination; and (ii) a wrist-worn microcontroller with a second IMU, which simulates a smartwatch and provides

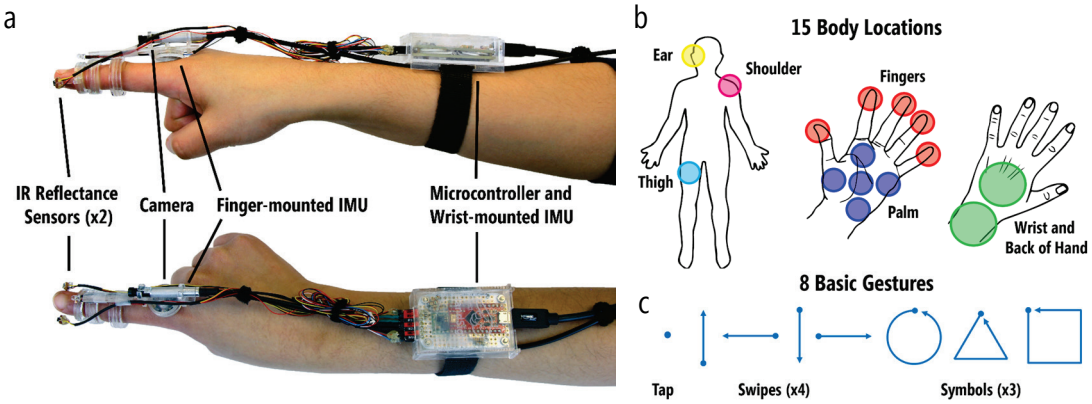


Figure 7.2: (a) TouchCam Offline showing the finger and wrist-worn sensors and microcontroller. (b) Fifteen fine-grained body locations (individual circles) within six coarse-grained locations (denoted by color), and (c) eight basic gestures.

additional sensing. The finger-based sensors are mounted on three laser-cut rings and positioned to avoid impeding the user's sense of touch, which is particularly important for users with visual impairments. See Figure 7.2a.

Infrared Reflectance Sensors. The two IR sensors[25] (each 2.9 mm × 3.6 mm × 1.7 mm) have a sensing range of ~2–10 mm and are used to detect touch events and to aid in recognizing gestures. Unlike Magic Finger [228], which places optical sensors directly on the pad of the finger, we mount the sensors on the sides of the front-most ring, approximately 5 mm from the fingertip to avoid interfering with tactile sensitivity.

Camera Sensor. A small (6 mm diameter) CMOS camera[26] is mounted atop the user's index finger, providing 640 × 640 px images at up to 90 fps with a 30° diagonal field of view (FOV). We use grayscale images from the camera to classify the touch location, extracting both texture and 2D point features. We also estimate visual motion between video frames to assist in classifying gestures. The camera includes a manually adjustable lens with a focal distance that varies from 15 mm to ∞. Because distances shorter than 30–40 mm provide a very narrow range of focus, we positioned the front of the lens ~50 mm from the user's fingertip. This setup provides an effective FOV of ~27mm across the diagonal when the finger is touching a surface. To prevent lateral rotation and to fix the FOV center near the touch location, the camera is attached

[25] Fairchild Semiconductor QRE113GR
[26] Awaiba NanEye GS Idule Demo Kit

to two rings 15 mm apart. A bright LED (3 mm diameter, 6000 mcd, 45° angle) mounted below the camera illuminates the touch surface regardless of ambient lighting.

Inertial Measurement Units. Two IMUs[27] are mounted on the user's hand: one below the camera on the index finger and one on the wrist. We include two IMUs to examine the effect of sensor location on classification performance. The IMUs provide motion information at ~190 Hz, and each contains a three-axis accelerometer, gyroscope, and magnetometer. While the camera offers rich contextual information about a scene, its field of view and frame rate limit performance during quick motions. Therefore, the IMUs are our primary sensor for detecting motion and classifying gestures. The orientation of the gesturing finger and/or wrist may also be useful for distinguishing body locations (*e.g.*, ear vs. thigh) although this is posture dependent. The IMUs are calibrated to correct magnetic bias and to establish a stable orientation estimate (described in Section 7.1.2). Calibration consists of rotating the unit along each axis for a few seconds and is performed only once per session—although future explorations may require repeated calibration to ensure long-term stability.

Sensor Placement and Microcontroller. We designed custom laser-cut rings in multiple sizes (13–24 mm inner diameter in 0.5mm increments) with detachable sensors to fit each user. As shown in Figure 7.2a, the rings are worn on the index finger near the first and second joints. The IR and IMU sensors are controlled via a microcontroller[28] mounted on a Velcro wristband, and the camera and microcontroller

[27] Adafruit Flora LSM9DS0

[28] Sparkfun Arduino Pro Micro (5V/16MHz)

are connected to a desktop computer[29] via USB cables. All data is logged, timestamped, and analyzed *post hoc* on the desktop.

7.1.2 Input Recognition Algorithms

To recognize localized on-body input, we developed a four-stage approach: (i) touch segmentation; (ii) feature extraction; (iii) location classification; (iv) gesture classification. The two classification stages—location and gesture—are trained individually for each user and combine readings from multiple sensors for robustness. While the algorithms described next could be trained on any arbitrary set of locations and gestures, in our study, we evaluated six coarse-grained body locations (*fingers*, *palm*, *back of hand or wrist*, *ear*, *shoulder*, and *thigh)* with 15 fine-grained locations (*thumb/index/middle/ring/pinky finger*, *palm up/down/left/right/center*, *back of hand*, *outer wrist*, *ear*, *shoulder*, and *thigh*) and 8 basic gestures (*tap*, *swipe up/down/left/right*, *circle*, *triangle*, and *square*)—see Figures 7.2b and 7.2c. These locations are visually distinctive and can be located easily even without sight, and the gestures are simple shapes that can be drawn with a single touch down/up event.

Stage I: Touch Segmentation. Our input recognition algorithms receive a sensor stream consisting of video, IMU, and IR data. We segment this input stream by detecting touch-down and touch-up events using the IR sensor readings, which represent distance from the touch surface (lower values are closer). While for real-world use, a segmentation approach would need to identify these touch events within a

[29] Dell Precision Workstation, dual Intel Xeon CPU @2.1GHz, NVIDIA GeForce GTX 750Ti

continuous stream of data, to evaluate this initial prototype we made several assumptions to simplify the process (we eliminated these assumptions for the realtime prototype, described later). Based on experiments with pilot data, we developed a straightforward threshold-based approach using a variable threshold that was set to 90% of the maximum IR value observed across the input stream for each trial. Within a trial, a *touch-down* event is triggered when either of the two IR values crosses below the threshold, while a *touch-up* event is triggered when both cross above the threshold. To be conservative, we assume that each trial contains a single gesture and segment the entire gesture from the first *touch-down* event in the trial to the last *touch-up* event. We crop each input stream to include only the sensor readings and video frames that occurred between the touch-down and touch-up event timestamps.

Stage II: Feature Extraction. In Stage II, we extract static orientation and visual features for localization, and motion features for gesture classification. We describe each in turn below (see Table C.4 in Appendix C for more details).

Localization Features. To extract static features for localization, we first determine the video frame that has the maximum focus in the segmented sequence, since it is the most likely to contain recognizable visual features. We define focus as the total number of pixels extracted using a Canny edge detector [21] tuned with a small aperture ($\sigma = 3$) and relatively low thresholds ($T_1 = 100, T_2 = 50$) to detect fine details. While this approach does not account for all image quality problems—motion blur in particular can cause it to fail—it is highly efficient and, in general, detects a much greater number of edges for images that are in focus than for those that are not.

We verified this trend empirically using pilot data. We then extract several features for the selected video frame, which include: (i) raw IR sensor readings, (ii) the estimated IMU orientation, (iii) image texture features for coarse-grained classification, and (iv) 2D image keypoints for geometric verification to distinguish between locations with similar textures (*i.e.*, fingertips, palm locations, back of wrist or hand).

The orientation of each IMU is estimated by applying a Madgwick filter [124] on a sequence of raw accelerometer, magnetometer, and gyroscope readings resulting in a 4D orientation vector (*i.e.*, quaternion). The filter is a standard sequential optimization approach to estimating IMU orientations that is updated at each time step. Our initial calibration procedure includes briefly rotating the device in all directions so that the filter can converge to an accurate orientation estimate. The orientation estimate at the selected video frame is used as a 4-dimensional feature vector (W, X, Y, and Z) for each IMU and concatenated into an 8-dimensional vector when both IMUs are used.

The image-based features are extracted similarly to our prior work in the previous chapter: To represent image texture, we use a variant of local binary patterns (LBP) that is robust to changes in illumination and that achieves rotation invariance while exploiting the complementary nature of local spatial patterns and contrast information [62]. While we explored other common texture-based methods such as Gabor histograms [216] and wavelet principal components [44], we found that they offered negligible improvements over LBP despite their increased computational complexity. We extract uniform LBP patterns and local variance estimates from an image pyramid with eight scales to capture both fine and coarse texture information.

Specifically, we use $LBP_{12,2}^{riu2}/VAR_{12,2}$ with 14 uniform pattern bins and 16 variance bins as defined in [62]. These values are accumulated into a histogram with 224 bins for each scale, all concatenated to obtain a 1792-element feature vector. To resolve ambiguities and ensure geometric consistency, we extract custom keypoints at locations with a high Gabor filter response at two or more orientations, which tend to lie at the intersections of ridgelines or creases. This approach was inspired by [80]. We use the Gabor energy in a 16 × 16px neighborhood around the keypoint as a descriptor extracted at 18 orientations to ensure rotation invariance. See Chapter 6 for full details.

Motion Features. For gesture classification, we extract motion features from the sensor readings within the segmented timeframe (these are treated independently from the localization features). We use three standard signal preprocessing steps on the raw IMU and IR sensor readings: smoothing, normalization, and resampling. We first smooth the raw values using a Gaussian filter (σ =13, optimized based on pilot data) to reduce the effect of sensor noise and then normalize the smoothed sequence by subtracting its mean and dividing by its standard deviation. To obtain a fixed length sequence for robustness to variations in speed, we resample the sensor readings using linear interpolation at 50 equally spaced discrete time steps. These values, however, are still sensitive to small variations in speed and orientation. Thus, similar to [224], for each IMU and IR sensor we compute summary statistics for windows of 20 samples at 10-step increments (*i.e.,* four windows): mean, minimum, maximum, median, and absolute mean. Finally, for the 50 resampled accelerometer, magnetometer, and gyroscope readings, we compute *x*-*y*, *x*-*z*, and *y*-*z* correlations. The result is 639 features

for each IMU and 70 for each IR sensor, which we concatenate into a single feature vector to use when classifying gestures.

We also extract motion features from the video frames between the touch-down and touch-up events. Because we support touch-based gestures only on flat or nearly flat surfaces, it is sufficient to estimate a global 2D motion vector for each frame; we do so using a template-matching approach. First, we down-sample each image from 640 × 640px to 160 × 160px resolution for efficiency and noise robustness. Next, for each frame we extract a 40 × 40px region centered within the previous frame to use as a template, which we then match against the current frame using a sliding window to compute the normalized cross-correlation [115]. The position of the pixel with the highest cross-correlation value identifies the most likely displacement between frames, yielding a 2D motion vector estimate. Because images with higher contrast are more likely to yield reliable motion estimates, we weight each motion vector by an estimate of the frame's contrast (the image variance). As with the other motion features, we smooth the motion estimates by applying a moving average (window size = 10). We then re-sample 50 points from this sequence of motion vectors and compute summary statistics as with the IMU and IR sensor readings to obtain a fixed-length vector of 140 features for use in gesture classification.

Stage III: Localization. Once we have extracted localization and motion features, we begin independently classifying on-body locations (Stage III) and gestures (Stage IV). For localization, we rely primarily on static visual features from the camera with IMU orientations and IR reflectance values to resolve ambiguities.

Our image-based touch localization algorithms function identically to our prior work [202]. We use a two-level location classification hierarchy: first classifying the location as one of the six course-grained regions then refining that location estimate where possible to finer-grained regions. In our offline user study (Study I), coarse-grained regions include *fingers*, *palm*, *back of hand or wrist*, *ear*, *shoulder*, and *thigh* while fine-grained regions include specific fingertips, locations on the palm, and on the back of hand versus the wrist (Figure 7.2b). Some coarse-grained locations are not subdivided at this second level due to a lack of distinctive features—in the case of our study, the *ear*, *shoulder*, and *thigh* are not subdivided. We first classify the texture features into a coarse-grained location using an SVM[30] then perform template matching against only the stored templates from that location to estimate the fine-grained location. Finally, we perform geometric verification using the extracted 2D point features to ensure a correct match.

At both levels of the classification hierarchy, we resolve ambiguities using a sensor fusion approach. We combine predictions based on the static visual features from a video frame with predictions based on the IMU orientation and IR reflectance features with the same timestamp as that frame. Since the scales, lengths, and types of these feature vectors differ greatly, rather than concatenating the features for use with a single classifier we instead train a separate SVM with a Gaussian kernel on the non-visual features. To robustly combine the predictions from the two disparate localization

[30] Aforge.NET: http://www.aforgenet.com (used for all SVM and neural network classifiers)

classifiers (one for the camera features and one for the IR and IMU features), we first tune the SVMs to output normalized probability predictions for each class using Platt scaling, as is standard [165]. We concatenate these predictions into a single feature vector, which we then use to train a third sensor fusion classifier that automatically learns how to prioritize sensors based on prediction confidence and location class. Inspired by [38], we use a feedforward neural network for this sensor fusion classifier. Our network has one fully connected hidden layer for flexibility of functional representation, and a softmax output layer for multiclass output; it is trained using resilient backpropagation [129]. The final output of our classification process is a combined location prediction from the six coarse-grained and fifteen fine-grained classes with approximate likelihoods for each class (sorted from most to least likely).

Stage IV: Gesture Classification. Gesture classification is performed independent of localization using an additional SVM. As in texture classification, SVMs are commonly used for classifying gestures because they are robust and efficient for problems with high dimensionality. We use a linear kernel with feature weights that were optimized for performance across all participants. For the evaluation presented in Section 7.1.4, we trained an SVM to classify the following gestures as shown in Figure 7.2c: *tap*, *swipe up*, *swipe down*, *swipe left*, *swipe right*, *circle*, *triangle*, and *square*.

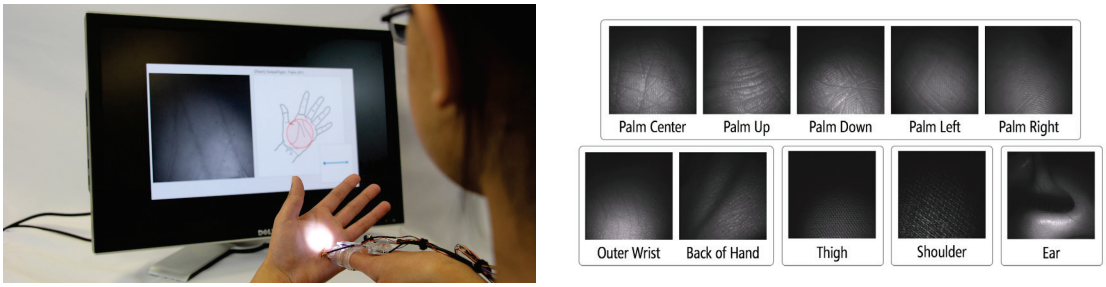
7.1.3 Study I: Data Collection and Dataset for Offline Experiments

To evaluate our initial prototype and algorithms, we performed offline experiments using data collected from twenty-four participants. Each participant performed a series

of location-specific on-body input tasks with our prototype system. We were specifically interested in investigating our first two research questions enumerated in the Introduction: (i) How accurately can we recognize location-specific on-body gestures with a finger-worn camera and auxiliary sensors (IMU, IR)? (ii) Which body locations and gestures can be recognized most reliably using our approach?

Participants. Twenty-four right-handed participants (16 female) were recruited via campus e-mail lists and word of mouth. Their average age was 28.9 (*SD*=7.95, *range*=19–51). All participants had normal vision as the goal of this study was to assess our algorithms and not issues related to usability or accessibility. Participants were compensated $25 for their time.

Data Collection Apparatus. During data collection, participants wore the TouchCam Offline prototype. As described in Section 7.1.1, we selected ring sizes to fit the participant's finger and adjusted positioning to ensure a consistent sensor range. A custom application written in C# displayed visual task prompts and a live feed from the finger-worn camera to assist with framing the target locations (Figure 7.3a). All



(a) Following on-screen data collection protocol **(b) Example skin images from Study I**

Figure 7.3: (a) Data collection setup showing our prototype, location and gesture instructions, and camera video feed. (b) Example skin-surface images recorded by our finger-mounted camera (fingerprint images omitted to protect our participants' privacy).

IMU and IR sensor readings and camera video frames were logged with timestamps along with ground-truth touch location and gesture labels for each trial.

Procedure. The procedure lasted up to 90 minutes. After a brief demographic questionnaire and setup period (*i.e.*, selecting rings, putting on the prototype), participants completed the following tasks, in order:

Location-specific touches. Participants touched and held their finger in place at 15 locations (Figure 7.2b) with each location prompted visually on a monitor (Figure 7.3a). After confirming the location and image quality, the experimenter logged the current location (*e.g.,* timestamp, location label) and triggered the start of the next trial. Participants completed 10 blocks of trials, where each block consisted of a different random permutation of the 15 locations (150 trials in total). In total, this dataset includes 3600 *location-specific touches* across all participants. Example images are shown in Figure 7.3b.

Location-specific gestures. Participants performed the eight basic gestures: *tap*, *swipe up*, *swipe down*, *swipe left*, *swipe right*, *circle*, *triangle*, and *square* (Figure 7.2c) at three body locations: the *palm*, *wrist*, and *thigh*. These locations were selected from the 15 locations in the first task because they are easy to access, unobtrusive, and have a relatively large input area thus allowing for more complex gestures. As with the first task, participants completed 10 blocks of trials, where each block consisted of a different random permutation of the 24 gesture and location combinations (240 trials in total). This dataset includes 5,760 *location-specific gestures* across all participants.

7.1.4 Study I: Offline Experiments and Results

To investigate the accuracy of our location and gesture classification algorithms, we performed a series of offline experiments using the gathered data. Below, we evaluate coarse-grained localization, fine-grained localization, and location-specific gesture classification as well as the effect of each sensor on performance (*e.g.,* finger-worn *vs.* wrist-worn IMUs). We compare sensor combinations using paired t-tests and Holm-Bonferroni adjustments to protect against Type I error [77].

Training and Cross Validation. All of our experiments use leave-one-out cross validation and train and test on a single user's data. Specifically, each experiment uses all available data from a single participant for training the location and gesture classification SVMs with a single sample set aside for testing. The localization and gesture classifiers are trained independently. The experiment is repeated for each sample and averaged across all possible combinations.

Touch Localization. To examine the accuracy of our on-body localization algorithms, we used the location-specific touch dataset. Since our localization approach is hierarchical, we analyze performance at both the coarse-grained level (6 classes) and the fine-grained level (15 classes).

We first report primary localization results using all available sensor readings (*i.e.,* sensor fusion results). At the coarse-grained level, we achieve 98.0% (*SD*=2.3%) average accuracy. This is reduced to 88.7% (*SD*=7.0%) at the fine-grained level. Table 7.1 shows the accuracy breakdown by class. The worst performing coarse-grained classes were the wrist/hand and ear, both at 93.8%, possibly due to their highly variable

	Palm	Fingers	Wrist/Hand	Ear	Shoulder	Thigh
Palm	99.1%	0.5%	0.4%			0.1%
Fingers	0.3%	99.6%				0.1%
Wrist/Hand	5.0%	0.2%	93.8%	0.2%	0.6%	0.2%
Ear	4.2%	0.4%	1.2%	93.8%	0.4%	
Shoulder	0.8%		0.4%		98.8%	
Thigh	2.3%		0.4%			97.3%

(a) Coarse-grained Accuracy

	Up	Down	Left	Right	Center
Palm	84.6%	78.5%	85.0%	83.1%	91.5%

	Thumb	Index	Middle	Ring	Pinky
Fingers	93.1%	85.4%	81.5%	88.1%	91.9%

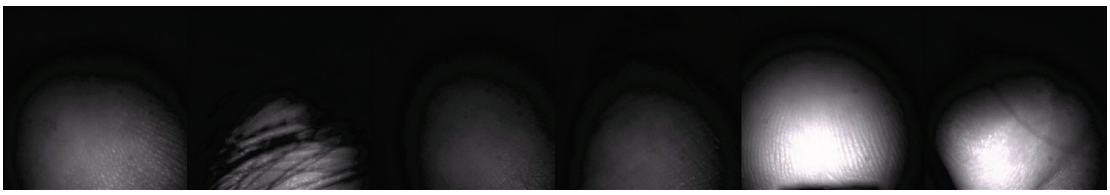
Outer Wrist	Back of Hand	Ear	Shoulder	Thigh
87.3%	88.8%	93.8%	98.8%	97.3%

(b) Fine-grained Accuracy

Table 7.1: Classification percentages averaged across 10 trials and 24 participants. (a) Accuracy for the six coarse-grained classes. Each cell indicates the percentage of images assigned to a predicted class (column) for each actual class (row); empty cells indicate 0%. (b) Accuracy for the 15 fine-grained classes, grouped by corresponding coarse-grained class.

appearance and fewer distinctive visual features. In contrast, the fingers and palm perform best at 99.6% and 99.1% respectively although the individual fine-grained classification accuracies were lower. These results suggest that care must be taken in selecting body locations that are both visually distinctive and easy for participants to return to repeatedly. A qualitative analysis of our dataset revealed issues that account for some of the error: approximately 5% of the images gathered had focus, contrast, or illumination issues that interfered with extracting recognizable image features; see Figure 7.4. We took steps to mitigate these problems for the next TouchCam iteration.

To investigate the effect of each sensor on localization performance, we repeated the classification experiment with the sensors individually and in combination. As expected, the camera is by far the most accurate single sensor for classifying location, with a coarse-grained accuracy of 97.5% (*SD*=2.6%) followed by



Out of Focus **Too Dark or Poor Contrast** **Oversaturated**

Figure 7.4: Approximately 5% of the images we collected had poor focus, contrast, or illumination, preventing robust feature extraction. We adjusted the camera and LED to mitigate these issues for TouchCam Realtime.

the finger-based IMU at 75.6% (*SD*=11.6%). Notably, the camera is significantly better even compared to the 87.5% (*SD*=7.0%) accuracy of combining all other sensors ($p<0.001$, $t_{23}=7.12$, $d=1.92$). No significant differences were found between the camera alone or combined with other sensors, which suggests that the camera alone is sufficient for course-grained classification. At the fine-grained level, the camera is again the most accurate sensor (84.0%) even compared to all other sensors combined (52.9% accuracy; *SD*=12.0; $p<0.001$, $t_{23}=16.74$, $d=2.99$). But, unlike at the coarse-grained level, adding any of the other three sensors to the camera further increases accuracy, with the highest accuracy (88.7%) resulting from the combination of all available sensors.

Location-Specific Gesture Classification. To explore the possibility of supporting location-specific gestures, we conducted a classification experiment with the data from the location-specific gesture task (24 classes: 3 locations × 8 gestures). First, we classified the location using the image features from the camera (extracted from the video frame with maximal focus as described above). Since the location features from the IR and IMU sensors did not make a significant difference at the coarse-grained level, we omitted them here. Location accuracy for these three locations was 99.1% (*SD*=1.0%). Next, we classified the gesture using the motion features from all of the sensors (IMU, IR, and camera) achieving an accuracy of 96.6% (*SD*=2.6%).

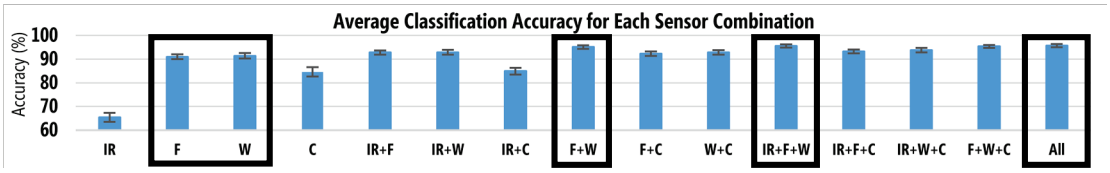


Figure 7.5: Mean classification accuracy using different sensor combinations to classify location-specific gestures. Boxes indicate the best sensor combinations as additional sensors are added, with each box significantly outperforming the last (from left to right). There was no significant difference between the finger- and wrist-mounted IMUs.

Finally, we combined the class predictions and calculated the overall location-specific gesture classification accuracy across all 24 classes, which was 95.7% (*SD*=3.2%).

As a secondary analysis, we again examined classification accuracy as a function of each sensor (Figure 7.5) but this time for the 24 location-specific gestures. In general, adding more sensors significantly improves classification accuracy, although as a practical matter the differences between the pair of IMUs and other more complex combinations are fairly small (see Appendix C for statistical comparisons).

Efficiency. For our initial prototype and algorithm development, our primary aim was to investigate the feasibility and accuracy of our approach rather than develop a realtime system. As such, our TouchCam Offline algorithms are slow. On our desktop computer (the Dell Precision Workstation described in Section 7.1.1), the image feature extraction and localization stages required, on average, two seconds *per frame* to process and classify an image. The most computationally demanding stage was geometric verification, which required approximately 243,000 feature comparisons on average. The other stages' computation times are comparatively negligible.

7.1.5 Summary of Study I Findings

Our results address our first two research questions demonstrating the feasibility of recognizing location-specific gestures using finger- and wrist-worn sensors. While our experiments show advantages with sensor fusion when classifying both location and gesture, the practical differences are relatively small suggesting that we can simplify our algorithms by using each sensor type for the task for which it is best suited (*i.e.*, IR

sensors for touch detection, camera for localization, and IMUs for gesture recognition). Individual accuracies per location suggest limits to the localization granularity of our algorithms, which performed well (≥98%) for coarse-grained locations but were less accurate (89%) for fine-grained locations. These results could likely be improved with a better camera (*e.g.,* higher resolution, autofocus) and with more complex finger/palm print recognition algorithms. However, the high accuracy during our location-specific gesture experiment (96%) suggests that such steps may not be necessary for us to begin investigating these interactions with visually impaired users. We built upon these findings to implement the next iteration of TouchCam, described below.

7.2 TouchCam Realtime: Improved Interactive Prototype

Based on our Study I findings, we designed *TouchCam Realtime*, a realtime version of our offline system with updated hardware and algorithms. We first describe key changes to improve robustness and enable realtime interactions (addressing RQ3) before validating the new classification algorithms using the Study I data.

7.2.1 Realtime Prototype Hardware

TouchCam Realtime's hardware (Figure 7.6) embeds all finger-mounted components in a single 3D-printed unit, which is attached to the user's finger by a pair of Velcro strips to allow greater freedom of motion than the rigid rings from the previous version. This updated design is more stable and durable. The camera and IR sensors are repositioned to capture more consistent images and improve the reliability of touch detection, respectively. Although Study I found an accuracy advantage when using two

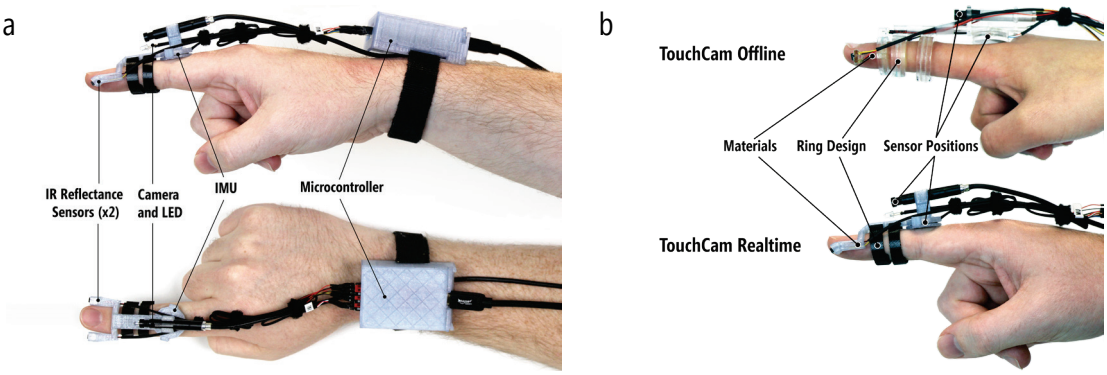


Figure 7.6: (a) TouchCam Realtime prototype showing the finger and wrist-worn sensors and wrist-worn microcontroller. (b) Comparison of TouchCam Offline and Realtime hardware.

IMUs (~4%), we decided to remove the wrist-mounted IMU to simplify our hardware and algorithms. We compensated for the potential drop in accuracy by doubling the remaining (finger-mounted) IMU's sampling rate. This change reduced the number of features used to classify gestures and the number of examples needed for training.

7.2.2 Realtime Input Recognition Algorithms

We made several changes to our input recognition algorithms to support realtime operation. First, we optimized our localization algorithms to run in realtime on a GPU and removed the computationally costly geometric verification step. Second, we updated the touch detection stage to support continuous use. Finally, we improved the gesture recognition stage, making it more robust to changes in orientation and pose.

Localization Algorithm Changes. As noted previously, our offline localization algorithms required up to two seconds per frame, primarily limited by geometric keypoint matching between image templates. Simply removing keypoint matching increased our frame rate from 0.5fps to ~18.5fps, but with a ~9% reduction in Study I's fine-grained localization accuracy. To address this loss, we made three

updates to our localization algorithms. First, we used an alternate LBP approach that better preserved spatial features [236], which increased the number of texture features per image from 1792 to 15,552. Second, we averaged class probability predictions across 20 video frames, a number selected after pilot tests to balance accuracy and latency. And third, we reduced the number of fine-grained locations, omitting the five fingertip locations evaluated in Study I. This decision was not solely due to algorithmic performance—the fingertips proved difficult for participants to capture reliably even with visual feedback due to the sensors' positioning and small field of view. Also, while the fingertips are convenient locations for static touch-based input, they are too small to easily support gestural input. Finally, we implemented parallel GPU versions of our algorithms, which further improved the average localization speed to 35.7 fps.

Touch Detection Algorithm Changes. To improve robustness and support continuous use, we made minor changes to the touch detection algorithms. We applied a moving average filter to the IR values to reduce sensor noise (*window size* = 50ms), and triggered *touch-down* and *touch-up* events when the sensors crossed a fixed threshold that was the same across all users rather than derived per gesture as with the offline system. This threshold was fixed at 90% of the maximum possible value the sensor could register, which we determined empirically to be robust to changes in ambient lighting and to work well for skin and clothing surfaces. To ensure that we captured the full gesture (and to support double-taps), we placed a delay of 100ms on the *touch-up* event and canceled it if the user touched down again within that period.

Gesture Recognition Algorithm Changes. Lastly, we made improvements to the gesture recognition algorithms. To compensate for variations in orientation and pose when performing gestures, we first rotated the IMU sensor readings relative to the estimated orientation at the start of the gesture (the *touch-down* event). We discarded the magnetometer readings after this step since they were overly sensitive to orientation and location. These changes allowed us to build a pre-trained cross-user gesture classifier with 1,720 samples in place of the individual classifiers used in Study I.

7.2.3 Validation of Realtime Algorithms

To test our updated algorithms and establish a performance benchmark for our realtime system, we conducted classification experiments on the data gathered during Study I. The average 10-fold cross-validation accuracy on the location-specific touches dataset was 97.5% (*SD*=2.4%) at the coarse-grained level (6 classes) and 84.5% (*SD*=8.2%) at the fine-grained level (15 classes), which is nearly identical to the TouchCam Offline system—see Table 7.2. The five finger locations were most impacted by the removal of the geometric verification step. Localization accuracy on the location-specific gestures dataset remains similarly high at 98.6%. As mentioned above, efficiency increased considerably: from 0.5fps to 35.7fps (a ~70x speedup).

	Palm	Fingers	Wrist/Hand	Ear	Shoulder	Thigh
Palm	98.5%	0.8%	0.5%	0.1%		0.1%
Fingers	0.3%	99.7%				
Wrist/Hand	4.0%	0.4%	95.4%			0.2%
Ear	5.4%	2.1%		92.1%	0.4%	
Shoulder	1.7%	0.4%	2.1%		95.4%	0.4%
Thigh	1.7%		2.1%	0.4%		95.8%

(a) Coarse-grained Accuracy

	Up	Down	Left	Right	Center
Palm	83.8%	82.5%	80.8%	85.4%	89.6%

	Thumb	Index	Middle	Ring	Pinky
Fingers	92.1%	71.3%	71.3%	73.8%	79.6%

Outer Wrist	Back of Hand	Ear	Shoulder	Thigh
87.5%	85.8%	92.1%	95.4%	95.8%

(b) Fine-grained Accuracy

Table 7.2: TouchCam Realtime performance on Study I dataset. (a) Coarse-grained classification averaged across 10 trials and 24 participants. Each cell indicates the percentage of images assigned to a predicted class (column) for each actual class (row). (b) Fine-grained classification averaged across the corresponding coarse-grained classes.

7.3 Study II: Realtime Evaluation with Visually Impaired Participants

To assess the performance and accessibility of TouchCam Realtime under more realistic conditions and with our target population (RQ4), we conducted a second study. We recruited 12 blind and visually impaired participants who performed common interactions with TouchCam such as checking the time or reading text messages. We focus primarily on issues impacting the accuracy and usability of our system (see [156] for more about the interaction designs and participant feedback).

7.3.1 Study II: Method

Participants completed an adaptive calibration procedure for training and then used TouchCam Realtime to perform tasks using three on-body interaction techniques.

Participants. Twelve participants (7 female, 5 male) were recruited through email lists, local organizations for people with visual impairments, and word of mouth. Nine participants were blind and three had low vision. The average age was 46.2 years old (*SD*=12.0, *range*=29–65). All participants were smartphone users (11 iPhone, 1 Android) and all reported using a screenreader either "all" or "most" of the time. Participants were compensated $60 for time and travel.

Apparatus. Throughout the study, participants wore the TouchCam Realtime prototype on their dominant hand. We assisted participants with putting on the ring and wristband and adjusted positioning to ensure consistent sensor readings. A custom C# application controlled a semi-automated training process, provided audio and

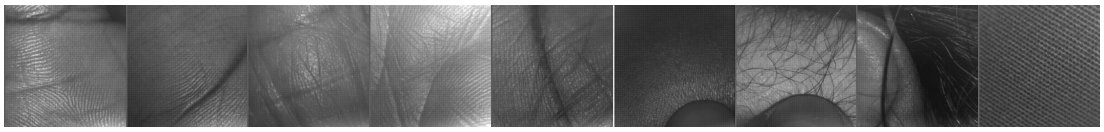


Figure 7.7: Sample image data from the nine locations collected with TouchCam Realtime. All images were selected from different participants.

synthesized speech cues during the tasks, and displayed a camera and sensor view for the researcher to ensure correct positioning. All sensor readings and video frames were logged with timestamps.

Location and Gestures. As described in Section 7.2.2, we refined the locations and gestures for Study II based on observations made during Study I. We reduced the coarse-grain set from 6 to 4 locations and the fine-grain set from 15 to 9 locations. Specifically, we replaced the back-of-the-hand location with the inner wrist due to inter-class similarity with the outer wrist, removed the shoulder location for ergonomic reasons, and removed the five finger locations because without 2D keypoint matching and geometric verification, classification accuracy for this region was considerably lower. The updated set of locations included: the palm (up, down, left, right, and center), the wrist (inner and outer), the thigh, and the ear (Figure 7.7).

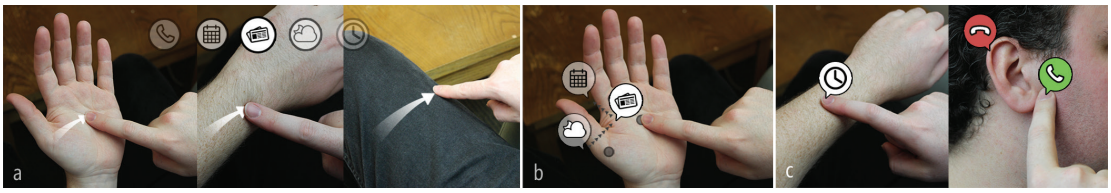
While Study I showed that TouchCam can support a variety of touch-based gestures, for Study II we specifically modeled our interactions after Apple's VoiceOver[31] and Google's TalkBack[32]—two popular gesture-based mobile screenreaders for non-visual use. In total, we support 6 gestures, including: *swipe left*

[31] http://www.apple.com/accessibility/ios/voiceover/
[32] https://play.google.com/store/apps/details?id=com.google.android.marvin.talkback

or *swipe right* to move between menu items and *double-tap* to select an item. We also included a *single-tap* gesture to repeat a voice prompt, a *swipe-down* gesture to go to the previous menu, and a *tap-and-hold* gesture to select location-specific items. The tap-and-hold gesture was recognized by an 800ms timeout after the *touch-down* event while the other gestures were recognized using a pre-trained SVM classifier (as described in Section 7.1.2). These gestures can be performed at any body location.

Training Procedure. To limit the amount of time needed to train our system, we implemented an adaptive training procedure inspired by boosting [49]. After capturing a single image of each of the nine locations for initialization, participants then moved their finger around each location in a fixed order as the system continuously classified the video frames. Whenever a video frame was misclassified, that frame and the current location label were saved, and the classifiers were retrained. This semi-automated training continued until convergence (*i.e.*, until the researcher determined that the automated system was performing well). After training all locations, at least one additional round was necessary to ensure that new image samples did not negatively affect performance. We found that the initial training images plus two rounds of semi-automated training were sufficient for most users, which took roughly 15-20 minutes and resulted in an average of 13 training examples per location (*SD*=4.5; *range*=5-24).

Procedure. The study procedure lasted up to two hours and consisted of: (i) an interview about mobile and wearable device usage including thoughts about on-body interaction (~20 minutes); (ii) system calibration and training (~30 minutes); (iii) using



(a) Location-independent gestures (*LI*) (b) Location-specific palm gestures (LS_{palm}) (c) Location-specific body gestures (LS_{body})

Figure 7.8: Three on-body interaction techniques: (a) for *LI*, users swipe left/right anywhere on the body to select an application. For (b) and (c), users select an application by double tapping on a specific location on their palm (LS_{palm}) or body (LS_{body}).

TouchCam with three interaction techniques (~10 minutes each); and (iv) a post-study questionnaire (~15 minutes). For (iii), the VoiceOver-like interaction techniques were presented in a fully counterbalanced order. Each interaction technique supported the same set of applications and menu items accessed through a two-level hierarchical menu. The top-level menu had five applications (*Clock*, *Daily Summary*, *Notifications*, *Health and Activities*, and *Voice Input*), which were selected by double tapping. Once selected, each application had 3-4 submenu items except for *Voice Input*, which had no submenu. The three interaction techniques are described below (see also: Figure 7.8 and the supplementary video at https://youtu.be/VREiWI_38BQ):

1. *Location-independent gestures (LI)*. Users swiped left or right anywhere to select an application.

2. *Location-specific palm gestures (LS_{palm})*. Top-level applications were mapped to five different locations on the palm. Users pointed directly to a location to select that application or searched for an item by sliding their finger between locations (similar to VoiceOver).

3. *Location-specific body gestures (LS_{body}).* Functioned similarly to LS_{palm} but mapped the applications to five different locations on the body rather than just the palm. We attempted to use intuitive mappings. For example, tapping the outer wrist for Clock and the ear for Voice Input. The other mappings were: the palm for Notifications, the inner wrist for Daily Summary, and the thigh for Health and Activities.

After activating an application, navigation of the submenus was identical across all three interaction techniques, using swipes left and right to select an item and a double-tap to activate it. For each of these interaction techniques, participants were instructed to complete the same set of 10 tasks in a random order. After an automated voice prompt said "begin," a task consisted of selecting an application, opening its submenu, and then selecting and activating a specific menu item (*e.g.*, "open the *Alarm* item under the *Clock* menu"). After the correct menu item had been activated by double-tapping, an automated voice prompt said "task complete," and participants proceeded to the next task.

The session concluded with open-ended questions about the participant's experience using TouchCam Realtime and the three interaction techniques.

Data and Analysis. Throughout the study, we logged all sensor readings, the location and gesture classifications, and event occurrences (*e.g.,* task start/end, menu navigation). We analyze performance in terms of classification accuracy, as well as qualitative metrics of robustness and usability for the three on-body interaction

techniques that we tested. We also describe qualitative reactions and subjective preferences based on the interviews and questionnaires.

7.3.2 Study II: Experiments and Results

To evaluate TouchCam's realtime performance and usability with our blind and low-vision users, we observed participants' behavior during the study and analyzed subjective feedback about our system. We also conducted offline experiments as with Study I, focusing on the sensor data gathered during the training phase of the study (rather than later data, which was unlabeled). Below, we summarize the details of our experiments and findings.

General Observations and Reactions. All twelve participants successfully used TouchCam Realtime to complete tasks with each of the three interaction techniques. In the pre-study interview, most participants (N=9) reacted positively toward the idea of on-body interaction citing quick and easy access (N=7), the ability to map specific tasks to different body locations (N=6), reducing the number of devices to be carried (N=6), and not needing to hold a phone in hand, thus avoiding the risk of theft or damage and potentially freeing that hand for other tasks (N=4).

Participants reacted similarly after using the TouchCam prototype. Preferences were split between the three interaction techniques. Participants appreciated the low learning curve and flexible input location of the *LI* interface, which supported simple swipe and tap gestures anywhere on the body, while the location-specific LS_{palm} and LS_{body} interfaces offered quicker and more direct selections once the location mappings

were learned. Some participants preferred the proximity of locations for LS_{palm} because it enabled easy exploration and minimal movement, while others liked the more intuitive location mappings of LS_{body}. Key concerns included TouchCam's large physical size, the occasional difficulty with the LS_{palm} interface due to its lower fine-grained accuracy, and the social acceptability of using LS_{body} in public (*e.g.,* touching an ear may draw unwanted attention to the device). See [156] for a more thorough examination of qualitative reactions to our system.

Localization Accuracy. To assess TouchCam's localization accuracy and robustness for visually impaired users, we analyzed the data gathered during the *training* phase of the study. We first conducted a leave-one-out cross-validation experiment using the recorded training samples for each participant (similar to Study I). This resulted in an average accuracy of 91.2% (*SD*=3.5%) at the coarse-grained level and 76.3% (*SD*=76.3%) at the fine-grained level, which is a reduction in performance compared to Section 7.2.3. This decrease, however, is reflective of our adaptive training procedure: since new samples are added only when misclassified using the current SVM, we would naturally expect lower performance when removing even a single sample for cross-validation.

Thus, we conducted an additional experiment using the full training set and classified other video frames from the training session (*i.e.*, those recorded between the stored training samples). Here, the accuracy increases to 94.2% (*SD*=5.0%) and 81.3% (*SD*=6.6%) respectively. These latter numbers better reflect actual usage performance since we could not reliably measure ground truth during the actual user study (*i.e.,*

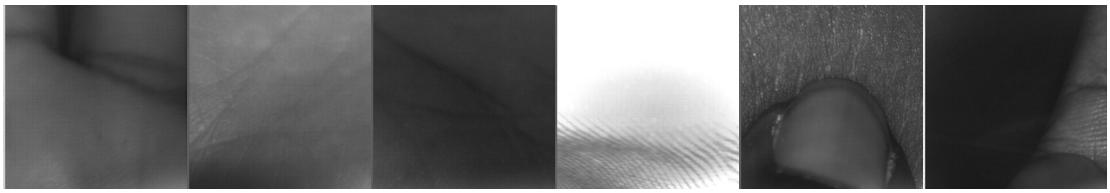


Figure 7.9: Some images captured during Study II were of poor quality due to the highlighted reasons. Despite these issues, performance remained adequate for participants to complete our specified tasks.

when participants were using TouchCam with the three interaction techniques). We note that although performance should be improved in future work (see Discussion), these results were sufficient for using and evaluating TouchCam with our participants.

Robustness. To investigate this reduction in performance in more detail, we performed a manual inspection of the 1,380 training images across the 12 participants using a custom image reviewing tool. While the severity of the problems varied widely, 22.2% of the images had some issue that could interfere with reliable classification (Figure 9), including: poor focus (13.6%), insufficient illumination (5.4%), poor contrast (4.3%), or oversaturation (0.8%). In addition, 3.2% of the images did not capture the target location due to the offset between the participant's touch location and the center of the camera's field of view, and in 0.6% of the images the participant's finger filled a large portion of the field of view, reducing the number of pixels available for identifying the target location. We further discuss robustness in the Discussion.

7.3.3 Summary of Study II Findings

Our findings validate TouchCam Realtime's performance with our target population and demonstrate three possible on-body interaction techniques that our approach can support. Participants successfully performed several simple input tasks with our

system, and their comments highlight positive reactions to on-body input as well as tradeoffs between the three interaction techniques. These tradeoffs reflect both TouchCam's performance (*e.g.*, LS_{palm} was least accurate due to its reliance on fine-grained localization) and broader design implications (*e.g.*, user preferences for flexibility of input location, learning curve, and social acceptability). Our findings also highlight obstacles to robust on-body input recognition, especially for visually impaired users who cannot rely on visual cues.

7.4 Discussion

While prior work has explored preliminary issues related to the design of on-body interfaces for visually impaired users [64,153], TouchCam is the first realtime wearable on-body input system designed for and evaluated with this population. Moreover, our work contributes the first real-time system for localizing skin images, and the first to explore location-specific touch-based gestures at a wide set of body locations. Below we discuss TouchCam's performance and usability and provide recommendations for future on-body input systems to support users with visual impairments.

7.4.1 Robust On-Body Input Detection Using Sensors on the Gesturing Finger and Wrist

Because TouchCam's sensors move with the gesturing finger, they can support touch input at a variety of body (and non-body) locations without requiring additional instrumentation. This feature allows greater input flexibility than most other on-body input approaches (*e.g.*, compared to *ViBand* [110] or Touché [184]) and means that the

user is also less likely to encounter issues with camera framing or occlusion—problems that are common for VI users when they use camera-based systems [3]. Although we did not examine non-body interactions in our work, TouchCam should support location-specific gestures at any surface with visually distinctive features.

Our results demonstrate the feasibility of a computer-vision driven finger-worn camera approach for on-body input; however, we also encountered obstacles that limit TouchCam's accuracy and precision. Because of the camera's size and positioning, image quality was variable. A high percentage (22.2%) of the training images gathered during Study II were out of focus, low contrast, or poorly illuminated, and in some images the target location was not visible due to the offset between the participant's touch location and the center of the camera's field of view. These usage issues appeared to have a greater impact on performance than other potential factors such as ambient lighting, skin tone, age, or hand size, although future work should investigate these possibilities in greater detail. Improved camera hardware could help address some problems—for example, autofocus functionality would help ensure sharp focus across changes in camera distance or perspective and a wider-angle lens would provide additional contextual information to aid classification. Audio feedback that notifies users when there is a problem and helps them learn how to use the system, as provided by assistive devices for reading such as *KNFB Reader*[33] or *OrCam*[34], could also be helpful. Finally, future work should explore hybrid sensing approaches that combine a

[33] http://www.knfbreader.com/
[34] http://www.orcam.com/

finger-mounted camera with additional body-worn sensors on the head or chest, which could provide additional contextual information and assist with localization.

7.4.2 An Expanded On-Body Input Vocabulary

As mentioned above, our work introduces new types of on-body interactions that other systems cannot readily support without additional instrumentation. For example, the fixed sensors used by ViBand's smartwatch platform limit interactions to a relatively small area on the hand and arm [110] while Touché requires modification of the target interaction surface and cannot detect gestural input [184]. In contrast, TouchCam can recognize location-specific gestures at several body locations, potentially allowing for intuitive context-specific input (*e.g.*, tapping the wrist to check the time) and supporting a high degree of flexibility and customization.

Participants identified tradeoffs between our three proof-of-concept interface designs, which should be considered when designing on-body interfaces to strike a balance between speed, accuracy, and learnability. Location-independent gestures (*LI*), which allow navigation using swipe gestures anywhere on the body, are easy to understand and learn, do not require individual calibration, and enable flexible input as needed for different situations (*e.g.*, sitting at home *vs.* walking while holding a cane). Location-specific gestures (LS_{palm} and LS_{body}), where the user can directly select an application or menu item by touching a specific location, are potentially quicker once the location mappings have been learned and can also support intuitive context-specific gestures as mentioned above. The palm-only version (LS_{palm}), with its high touch

sensitivity and close proximity between mapped locations, could enable faster and more discrete input. Compared with the other two versions, it also more readily supports "touch and explore" functionality that could help participants learn the location mappings more quickly. However, in our experiments LS_{palm} was less accurate than the other two because of inter-class similarity between palm locations and thus required participants to more carefully position their hand and fingers.

This expanded input vocabulary and flexibility of input locations may come at a cost, at least in the current iteration of TouchCam. While our prior work [202] suggested that we should be able to support precise localization on the palm and fingers using their rich visual features, our findings in this work highlight difficulty with robustly recognizing fine-grained locations. Future work should investigate ways to more reliably recognize fine locations, ideally with greater granularity than tested in our studies (*e.g.*, more than five palm locations), and recognizing touch input at two or more locations simultaneously (*e.g.*, using multiple finger-worn sensors) to support multi-touch gestures. In particular, future work should investigate how to extend our approach to support precise 2D localization (*e.g.*, as with OmniTouch [70] or CyclopsRing [25]). These goals may be possible with the aid of additional sensors (*e.g.*, a body-mounted camera) or with more efficient fingerprint and palmprint recognition algorithms that can support real-time interactions.

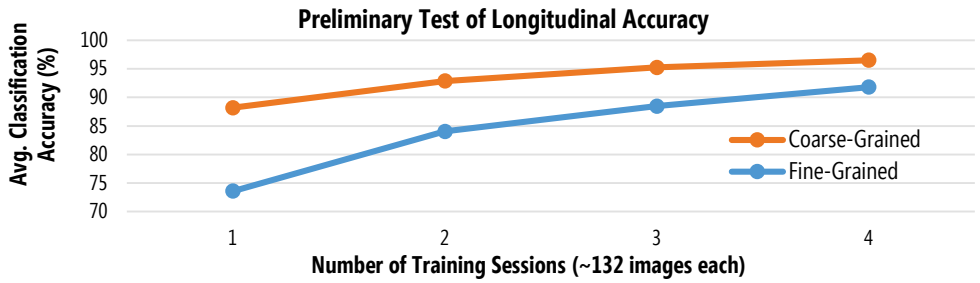


Figure 7.10: Classification accuracy across multiple sessions. In general, accuracy increases with more training sessions, suggesting that recalibration may initially be necessary, but that accuracy will eventually converge.

7.4.3 Training and Calibration

While TouchCam's gesture recognition algorithms are robust enough to allow for a shared classifier that works across users, its localization algorithms rely on unique skin and clothing features and must be individually calibrated for each user. This requirement raises two concerns: (i) the time needed to complete the individual training procedure, and (ii) the stability and robustness of the classifiers over time as the system shifts position and the user's body appearance varies (*e.g.*, due to changing moisture levels or clothing). We took steps to address the first concern in Study II by introducing our automated training procedure, which took about 15-20 minutes for a new user compared to 30-45 minutes in Study I. However, this procedure will likely need to be simplified and further streamlined in future versions. One possibility would be to bootstrap the system using a large amount training data across multiple users, which could enable coarse-grained classification without individual training. Fine-grained accuracy could be improved over time by learning as the system is used.

As for the second concern, it is possible (even likely) that shifts in the sensor positions after calibration negatively impacted performance for some participants

during Study II. Long-term performance is a challenge for many on-body input systems, since they can be highly sensitive to sensor positioning and biometric changes [234]. To explore how accuracy is affected over time, we conducted a small additional study with data gathered across five identical sessions with a single user (the first author). The time between sessions varied from 15 minutes to 24 hours, with the sessions completed over a three-day period. The prototype was fully removed between each session. Classification accuracy was measured similarly to the other experiments described above, except previous session data was used for training and the current session for testing.

As expected, accuracy drops considerably when training on a single session and testing on another, from the 94.2% coarse-grained and 81.3% fine-grained numbers reported in Study II down to 88.2% and 73.6% on average respectively. However, combining training data across sessions improves accuracy reaching an average of 96.5% and 91.8% at the two levels for four training sessions (Figure 7.10). A larger longitudinal study will be necessary to determine how well these results extend to other users and to a longer period of time, but these results are promising.

7.4.4 Physical Design

We designed TouchCam to avoid interfering with the user's movements and sense of touch, but the system is still large and requires tethering to a desktop computer for fast processing. With further algorithmic optimizations and increases in mobile processing power, we ultimately envision a smaller, self-contained system that uses a smartwatch

for processing and power. Furthermore, our priority with the finger-worn components was to ensure robustness and durability during our experiments, but our design can be streamlined considerably using existing technology. For example, the 6mm diameter camera module[35] that we selected could be replaced with a much smaller 1mm unit from the same manufacturer[36], and the IMU components could be embedded more directly into the ring (while the board we used is 16mm in diameter, the IMU itself is only 4mm square). The IR reflectance sensors positioned near the tip of the user's finger could potentially be replaced with an alternative touch detection method that is less intrusive—for example, an IR depth sensor with a longer range. Further work is needed to explore how these design changes impact accuracy, robustness, and user perceptions.

7.4.5 Limitations

Our system design and studies had several limitations. The TouchCam camera required manually focus adjustments and its relatively narrow field of view resulted in an offset between what the user was touching and what was sensed—the latter was particularly problematic for small locations (*e.g.,* finger tips). Future work should explore auto-focusing camera hardware with wide angle lens. The data collected during Studies I and II was collected under controlled conditions. Moreover, while the visually impaired participants in Study II were able to use TouchCam to complete all of the specified

[35] Awaiba NanEye GS Idule Demo Kit
[36] Awaiba NanEye 2D Sensor

tasks, they occasionally needed multiple attempts to do so. Future work should explore more realistic and longitudinal usage.

7.5 Summary

We introduced and investigated TouchCam, a variant of HandSight with additional sensors that was designed to support input at a variety of body locations while mitigating camera framing issues that blind users often experience. Our design also enables new types of contextual gestures based on location. We evaluated two iterations of the TouchCam system in terms of accuracy and robustness, as well as usability for our target group of visually impaired users. Our findings not only highlight the feasibility of our approach—greater than 95% accuracy at detecting 24 location-specific gestures, and support for realtime interaction at approximately 35 frames per second—but also characterize tradeoffs in robustness and usability between different types of on-body input. Fine-grained input on the palm and fingers is desirable for efficient and discrete input, but these locations are more challenging to classify reliably due to their small size and similar visual features; in contrast, disparate body locations are easier to recognize and may enable more intuitive mappings between location and application, but may also be less efficient for a new user and potentially socially unacceptable. Location-specific gestures have the potential to support efficient interaction for expert users, flexible input locations depending on user preference or situation (*e.g.*, while walking with a cane *vs.* sitting at home), task-based interactions tied to intuitive locations, and relatively fine-grained input for body areas that have

distinctive visual features (*e.g.*, fingertips and palm). In future work, we plan to explore ways to improve robustness and evaluate our system's long-term performance during a longitudinal study.

Chapter 8: Identifying Clothing Colors and Patterns

To extend HandSight's functionality beyond accessing printed text materials and controlling mobile devices, we applied it to the task of identifying clothing colors and patterns. As discussed in Section 2.2.2, while color identification tools for users with visual impairments are widely available (*e.g.*, [17,59]), they do not identify visual patterns or allow users to quickly inspect multiple locations—both of which are important for recognizing clothing [20]. For more advanced clothing pattern identification, Yuan et al. [209,227,231] developed systems to identify 4 patterns and 11 colors in images captured with a mobile phone or head-mounted camera. Blind users responded positively to the system, although more detailed identification of colors (*e.g.*, "rose red") and additional clothing patterns were desired. The interaction was also inefficient, requiring the user to hold out the clothing in front of them and use speech input to individually capture each still image to be classified.

In contrast, our finger-mounted camera approach allows users to move their finger across an article of clothing, combining tactile information with continuous audio description of the fabric's appearance (Figure 8.1). Positioning the camera and light source on the user's finger for touch-based interactions also mitigates issues with distance and lighting that can impact the accuracy of existing color and texture recognizers. Our approach is similar to *Magic Finger* [228], which was not intended specifically for visually impaired users but which similarly used a finger-mounted

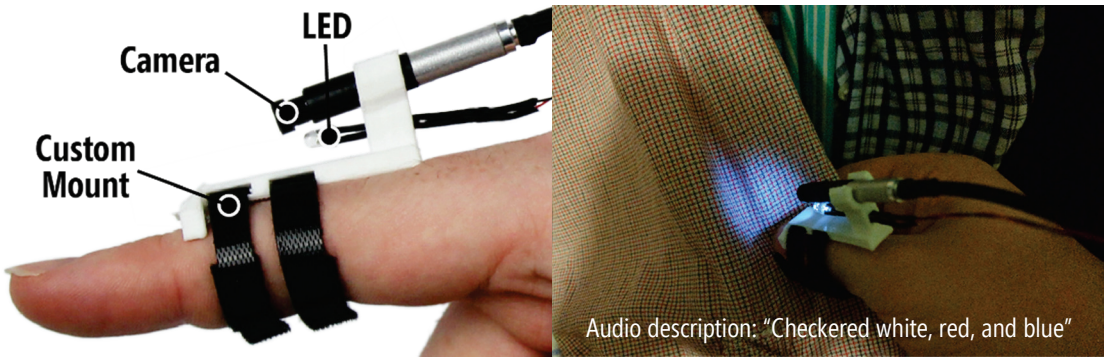
---

This chapter contains work published or scheduled to be published in the proceedings of the ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2017 [135]*, ASSETS 2018* [201]*).*

camera to classify touched surfaces including seven clothing textures. However, Magic Finger's classification approach was simplistic and would not scale well to a large database of textures. In contrast, we adapt state-of-the-art object classifiers for recognizing clothing fabric patterns using a combination of transfer learning and fine-tuning methods. Transfer learning is a machine learning technique used to adapt knowledge learned for one problem domain to another related domain [160], while fine-tuning is a process of refining a classifier's performance as additional data is gathered. To explore the feasibility of our approach and to test how reliably colors and visual patterns can be identified using close-up images from a finger-mounted camera, we collected two sets of fabric images and conducted offline experiments to assess performance. We achieve high accuracy at classifying fabric patterns (> 92%), and our findings suggest that HandSight could allow users to reliably identify unfamiliar patterns while shopping or train a specialized classifier for the articles of clothing in their closet. This chapter describes preliminary algorithmic work to assess feasibility, which has not yet been tested by visually impaired users; we close with a discussion of ongoing and future work toward implementing and evaluating a realtime interactive color and pattern recognition system.

8.1 Prototype System

To collect clothing images, we developed a simplified version of the HandSight hardware that included only the camera, LED, and a custom 3D-printed mount with Velcro straps (Figure 8.1). This is the same design used in Chapter 5 for supporting AR



(a) Close-up view of system (b) Identifying an article of clothing

Figure 8.1: Simplified prototype system for identifying colors and visual patterns

magnification. As shown in Figure 8.1, the system could be worn either on the index finger or thumb, and the position of the camera could be adjusted to allow us to test robustness and explore how much contextual information is necessary to reliably identify clothing patterns.

8.2 Initial Exploration: Visual Texture Classification

As an initial exploration, we tested an algorithmic approach based on the one described by Cimpoi *et al.* [32], which combines two complementary features commonly used for object recognition to achieve state-of-the-art texture classification performance. To examine how well this approach would extend to clothing images from a finger-mounted camera, we conducted a classification experiment on a small custom dataset.

8.2.1 Data Collection and Dataset

The results reported in [32] were promising but did not focus on clothing textures and used images from online sources that differed greatly from our target domain. To test how well the approach would extend to close-up images captured by a finger-mounted

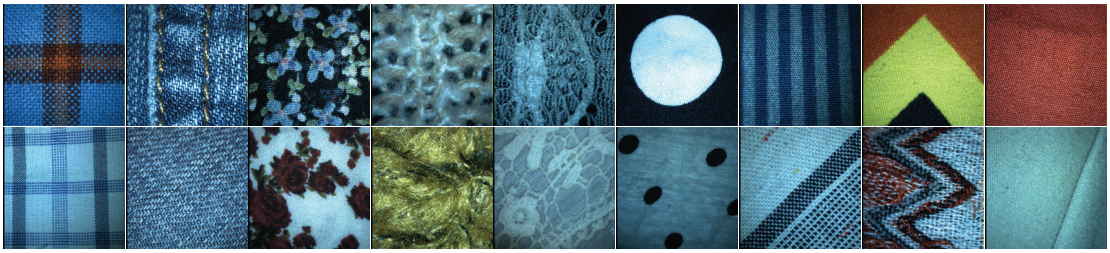


Figure 8.2: Examples of the 9 clothing textures included in our dataset. The numbers in parentheses indicate the quantity captured for each class. The full dataset can be downloaded at https://github.com/lstearns86/clothing-pattern-dataset.

camera, we collected a dataset of 520 images across 29 articles of clothing, which spanned 9 clothing texture categories (Figure 8.2). These categories are a subset of the 47 included in [32]; we eliminated categories that rarely describe clothing (*e.g.*, *bubbly*), combined those that are visually similar (*e.g.*, *striped*, *banded*), and added two new categories: *denim* and *none*. We controlled for and varied the distance (5cm *vs.* 12cm), rotation (0° *vs.* 45°), and perspective of the camera (90° *vs.* 45°), as well as the tension of the fabric (taut *vs.* hanging naturally); see Figure 8.3. The dataset was collected intermittently across controlled conditions by one person over three months.

8.2.2 Algorithms and Validation

To identify textures, we first compute deep convolutional activation features (DeCAF) using a pre-trained network. Our algorithms and experimental methods closely follow

Figure 8.3: We systematically varied distance, rotation, perspective, and fabric tension for each fabric sample collected using HandSight.

those used by Cimpoi *et al*. [32]. As in [32], we repurpose the *AlexNet* [105] image classifier for identifying textures by removing the last two softmax and fully connected layers and extracting the values in the exposed hidden layer as a feature vector. Second, we use scale-invariant feature transform (SIFT [122]) descriptors extracted densely at multiple scales. The SIFT descriptors are combined into a single feature vector based on their statistical distribution using the Improved Fisher Vector (IFV [164]) formulation. The result is two vectors of length 4,096 and 40,960 for DeCAF and IFV respectively, which are used as inputs (separately or concatenated together) to an SVM for classification.

To assess performance, we conducted a classification experiment computing accuracy as the number of test samples classified correctly. We also explored the effect of training set size to determine if a small user-gathered training set would be sufficient.

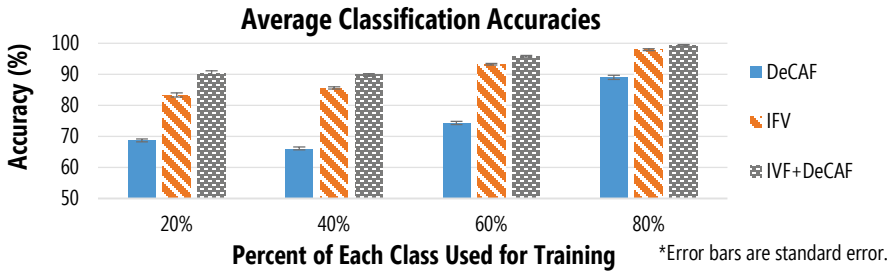


Figure 8.4: Accuracies using individual and combined features.

Figure 8.4 shows classification accuracy using DeCAF and IFV features separately and together as the percentage of data used for the training set increases from 20% to 80%, averaged across 40 random samples to reduce the effect of outliers. As with [32], in each case the combined result—that is, classification using a combined

vector of both DeCAF and IFV features concatenated together—is best, demonstrating the advantage of using complementary texture features. However, unlike [32], our DeCAF results are significantly lower than IFV. This is likely because DeCAF requires a large amount of training data to perform well while IFV does well even with the small amount that we provided.

8.3 An End-to-End Deep Learning Approach

While our initial exploration demonstrated the feasibility of our approach, the dataset was highly controlled, which risks overfitting, and the training process was not easily scalable. Additionally, the complementary feature approach was computationally demanding and did not take full advantage of modern deep learning techniques. To expand on that work, we built a larger and more varied dataset of images from online sources (Figure 8.5), which should allow our system to identify previously *unseen* fabrics—for example, to support shopping for new clothes. Unlike previous work using online images, we focused specifically on fabric images, and fine-tuned classifiers trained on the data with images collected using HandSight to improve performance in our target domain. To assess whether this Internet-based dataset can be used to identify patterns in images collected with our finger-mounted system, we adapted and fine-tuned a state-of-the-art deep neural network from an object classification problem and tested with the previously collected finger-mounted camera images.

Figure 8.5: Examples of the six classes in our fabric pattern dataset. The numbers in parentheses indicate the number of samples in each class (including augmentations). The full dataset can be downloaded at https://github.com/lstearns86/clothing-pattern-dataset.

8.3.1 Data Collection and Dataset

Existing texture datasets include textures that can easily be distinguished by touch or that are not relevant to clothing. For example, the *Describable Textures Dataset* [32] includes *braided* and *frilly*, and our initial dataset includes *denim*, *knitted*, and *lacelike*—which have unique textural patterns discoverable by touch. While automatic identification of these textures may be useful to avoid misclassifications, in general they are not necessary to assist blind users. Instead, we selected six common visual patterns that are difficult or impossible to distinguish by touch alone: *solid*, *striped*, *checkered*, *dotted*, *zigzag*, and *floral* (Figure 8.5).

To create our dataset, we added the word "fabric" after each class name and downloaded the top 1000 search results from Google Images using an open source utility[37]. After one person manually removed erroneous results and duplicates and

[37] Google Images Downloader, https://github.com/hardikvasa/google-images-download

cropped others as necessary (*e.g.*, to remove logos or background imagery), the dataset contained between 317 and 584 images per class (2764 images total). We augmented this data using a standard image synthesis process to increase the training set size and improve robustness [193], rotating each image in 30-degree increments and cropping the center at multiple scales (1–4 depending on the resolution of the original image), which resulted in 8232–17,304 samples per class or 77,052 images total.

8.3.2 Algorithms and Validation

To identify textures, we repurposed a state-of-the-art convolutional neural network model (ResNet-101 [73]) that was pre-trained on the *ImageNet* object dataset [180]. Using a standard transfer learning approach to avoid overfitting when insufficient data is available [41], we fixed all layers except for the final densely connected classification layer, and trained the weights for that layer using our dataset.

To ensure that each class was equally likely when training, we randomly sampled 6400 images from each class in the dataset for training and 1600 images for testing, discarding the rest. Classification accuracy on the test set was 91.7%, suggesting that this approach should work well in general. On our smaller finger-mounted camera clothing texture dataset (Section 8.2.1), which contained 400 images across the six classes, accuracy was 72.8%. Most errors were caused by confusion due to insufficient context or coarse threads (*e.g.*, Figure 8.6). For example, *zigzag* was the worst performing class, likely because the camera's proximity to the fabric obscured much of the pattern. Roughly 14.5% of images were also misclassified as *checkered* or

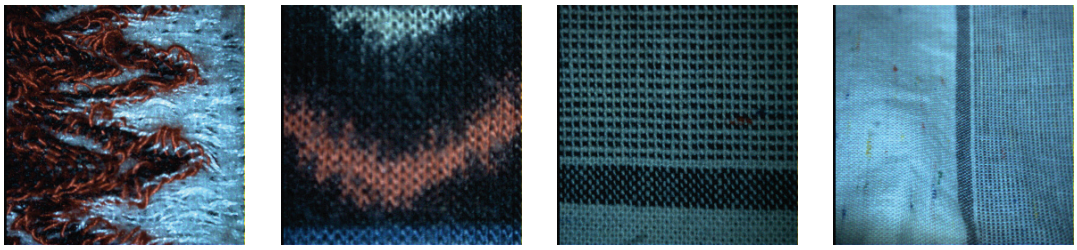


Figure 8.6: Example misclassifications (actual class → predicted class).

floral, most likely due to confusion from the coarse threads. Half of the images were captured with the finger-worn camera held 5cm from the fabric, while the other half were captured from a distance of 12cm; if only the latter images are considered (N=200), accuracy rises to 78.0%. Finally, fine-tuning the classifier using approximately half of the finger-camera images (N=36 per class) increases accuracy to 96.5%, suggesting additional images from the target domain will boost performance.

8.4 Discussion and Ongoing Work

Our preliminary results demonstrate the feasibility of recognizing clothing textures using close-up images from a finger-mounted camera, but many open questions remain. Here, we discuss issues relating to scalability and robustness, ongoing work on color recognition and description, and plans for a realtime implementation and user interface.

8.4.1 Scalability and Robustness of Pattern Recognition

Even with a small amount of training data across a variety of variables we achieve high classification accuracy (Section 8.2), suggesting that users could train a reliable personalized classifier by, for example, capturing images of the items in their closet. In our follow-up experiments with a larger dataset of Internet fabric images (Section 8.3),

pattern classification accuracy is similarly high when fine-tuned with images from a finger-mounted camera (97% vs 99%), but our end-to-end deep learning approach reduces training overhead and should be much more generalizable and robust. Still, open work remains to further improve robustness.

First, to mitigate errors caused by lack of context or distracting details (*e.g.*, coarse threads), the camera should likely be positioned farther back on the user's finger or wrist. This change would still allow users to easily query multiple locations and combine automated feedback with their own sense of touch. Second, as another avenue to improve robustness, future work should gather additional high-resolution images that show coarse fabric details, either using additional online sources (*e.g.*, querying "coarse fabric") or by collecting the images manually. This additional data could be used to teach the classifier to ignore details in the images that are unimportant for identifying the broader pattern.

8.4.2 Color Identification and Description

We focus primarily on visual *texture* classification since few researchers have attempted to make this information accessible for visually impaired users. However, our *color* recognition approach, described below, also has some advantages over existing solutions. For example, our wearable system uses touch-based interactions to constrain the camera's distance from the target surface and includes a bright LED to overpower the effects of ambient light, allowing for more consistent performance compared to existing color recognizers. Furthermore, we have explored two color

detection approaches that allow multiple colors to be detected simultaneously, and as mentioned previously our touch-based approach allows users to efficiently and interactively query multiple locations to build a mental model of how an article of clothing appears.

One common approach to identify the dominant colors in an image is k-means clustering. To allow for a variable number of colors, we use the "jump method" [205], which searches for the point at which adding additional clusters stops greatly reducing error. The k-means approach is straightforward and efficient, and preliminary tests on our datasets were promising. However, we also explored another approach based on superpixel segmentation [2], which groups neighboring pixels with similar colors together. Superpixel segmentation preserves spatial information, which could enable more reliable determination of which detected colors are salient when combined with the texture classification results (*e.g.*, omitting shadows and gaps between threads). To convey color information to users, we name each cluster center (or superpixel) using the XKCD color survey results [144] to provide commonly accepted names for 48 RGB color values. The level of detail is user-configurable, including the number and complexity of the names (*e.g.*, "green, purple, brown", or "lime green, lilac, beige"). Users can also identify multiple colors by moving their finger across the fabric. We did not evaluate accuracy, but if properly calibrated our finger-worn approach should mitigate issues with lighting and distance that impact existing color identifiers.

8.4.3 Realtime Implementation and User Interface

The color and pattern classification system runs at approximately four frames per second on a desktop computer[38]. Our current implementation tracks classification results for the most recent two seconds. To reduce noise from misclassifications, patterns are reported by majority vote, with unclear results labeled "unknown". Colors are only reported if they are named consistently across frames—for example, results of "blue and light blue" and "blue and gray" would be reported simply as "blue". Users can press a button to hear the most recent result via text to speech or hold for continuous updates. However, how best to convey complex color and pattern information to users is still an open question. Future work will also need to investigate performance and usability with visually impaired users and assess the potential benefits of our approach compared to existing aids.

8.5 Summary

We extend our finger-mounted camera system to recognize colors and visual patterns, which could allow visually impaired users to combine tactile information with a continuous audio description of a surface's appearance to, for example, obtain a better understanding of how an article of clothing appears. This work is preliminary and primarily algorithmic, but our results are promising: 99.1% accuracy when trained and tested on a small dataset (*e.g.*, as a personal classifier for the articles of clothing in a

[38] Dell Precision Workstation, dual Intel Xeon CPU @ 2.1 GHz, NVIDIA GeForce GTX 1080

user's closet, and 97.0% when trained on a much larger dataset of images from online sources, then fine-tuned and tested using images from a finger-mounted camera (*e.g.*, as a robust fabric pattern classifier that can identify previously unseen articles of clothing when shopping). We also discuss ongoing and future work on a complete system that can describe colors as well as patterns in realtime as users move their finger across a surface.

Chapter 9: Conclusion and Future Research Directions

The overarching goals of this dissertation were to improve information accessibility for people with visual impairments and to explore how touch-based access to non-tactile information can help users understand 2D surface content and appearance (*e.g.*, text, images, colors and patterns). To achieve these goals, we created the HandSight system, which augments the user's finger with interactive, real-time computer vision via a small wearable camera. Our work spanned two key application areas: (1) touch-based access to visual information in the physical world and (2) access to the digital information provided by computer and mobile devices through touch-based gestures. In this chapter, we summarize our high-level contributions before discussing broader implications and directions for future research.

9.1 Summary of Contributions

In this section, we restate the contributions listed in Chapter 1 and summarize how we achieved them. Our high-level contributions relate to the design, implementation, and evaluation of the HandSight system. We also summarize specific technical and design contributions across the three application areas for HandSight that we investigated.

9.1.1 The HandSight System

Our primary contribution is the development and iterative refinement of HandSight, a novel wearable system to assist people with visual impairments in their daily lives. Creating HandSight involved: (i) designing and implementing the physical hardware,

(ii) developing signal processing and computer vision algorithms, (iii) designing real-time auditory, haptic, and visual feedback that enables users with vision impairments to interpret surface content, and (iv) evaluating prototypes with visually impaired users to assess usability. We evaluated HandSight across a diverse set of tasks, providing both empirical evidence and qualitative user feedback that highlight tradeoffs when using finger-worn sensors to detect and recognize touched content or touch-based gestures in terms of the physical design, algorithmic complexity, and usability.

Physical design. A finger-worn design must by necessity use smaller components than designs worn on other body locations, restricting sensing fidelity, processing power, and battery life. For example, while HandSight was limited to 640×640 RGB images with a manual-focus lens due to size limitations, a larger camera could provide higher resolution and better image quality, autofocus capabilities, and new imaging features such as depth (*e.g.*, Omnitouch [67]) or hyperspectral imaging (*e.g.*, *HyperCam* [53]). Positioning the system on the finger also risks interfering with the user's freedom of movement and touch sensitivity. Because of these restrictions on physical size, weight, and positioning, additional care must be taken when designing the system to ensure durability, robustness, and social acceptability during daily use.

Algorithmic complexity. In terms of algorithmic complexity, a system using a finger-worn camera benefits from simplified processing when recognizing content at the tip of the user's finger and can use the full resolution of the camera for identification. Finger-worn sensors can also easily and directly detect touch events because of the sensors' proximity to the touched surface and can track relative motion

to recognize gestures—features we exploited to support flexible on-body input (Chapter 7). In contrast, to support touch-based interactions systems using cameras worn on the upper body or external to the user, the system must first locate the user's finger—which could be outside of the camera's field of view—and then determine if the user is touching a surface and identify the content that is beneath their finger using a small percentage of the camera's available resolution. However, a camera that is positioned farther away from the touch surface can capture additional contextual information for the system to use when interpreting content; a finger-worn camera has a more limited field of view and therefore the system must build up an internal representation of broader context as the user moves their finger across the surface. For example, HandSight can only capture a few words in each image when reading printed materials—a limitation shared with FingerReader [188–190] and other finger-wearables—while body-worn systems like OrCam [159] or handheld smartphone apps like KNFB Reader [98] can view and read full pages at once. Furthermore, performing global localization and motion tracking is much easier with an external view than with finger-based sensors, allowing systems using a more distant camera to, for example, more easily support location specific gestures on the user's palm or other input surface.

Usability implications. In terms of usability, finger-based sensing allows greater flexibility and a larger interaction space than other sensing approaches for touch-based interactions. Compared to non-wearables (*e.g.*, [91,184]), our approach is more portable and can support interactions on any surface by augmenting the user rather than the target content or input surface. Compared to a handheld smartphone or

dedicated device (*e.g.*, [99,243]), finger-worn sensors are more lightweight and hands-free, providing improved ergonomics and easier multitasking. And compared to cameras worn on the upper body (*e.g.*, [70,159]), our approach mitigates issues with framing the target content or gestures within the camera's field of view and allows interactions in a much wider area. When reading, our approach allows users to find a more comfortable position without needing to turn their head or body toward the target content. When performing gestural input, users do not need to perform the gestures directly in front of their body, allowing for more ergonomic and discrete input. A finger-worn approach also allows greater flexibility of input location for locations that upper-body cameras may not be able to sense (*e.g.*, the ear or thigh).

9.1.2 Technical and Design Contributions for Specific Applications

Here, we summarize the technical and design contributions that this dissertation makes across four specific application areas: helping blind users to read and explore printed materials, supporting augmented reality magnification for low vision users, recognizing location-specific on-body gestures to control computers and mobile devices, and identifying and describing clothing colors and visual patterns.

Reading and Exploring Printed Text. We first applied HandSight to helping blind users explore and read printed text materials. To assess how well users could trace and sequentially read lines of text by touch, we implemented haptic and auditory cues to guide the user's finger and systematically evaluated them across three user studies. We identified tradeoffs in terms of accuracy and user preference: audio may offer a

slight advantage to line-tracing accuracy and be more familiar to users but could also distract from the synthesized speech content; haptic uses a different sensory channel and potentially offers a clearer indication of direction but is less precise and may cause desensitization over time. Additionally, some participants in our studies appreciated the additional control over reading pace and the information about the positions of text blocks and images enabled by our design, which existing document scanner and screen reader approaches cannot easily provide. However, for common reading tasks participants preferred the experience provided by smartphone text recognition applications which, despite some difficulty aligning a document for capture, provided a faster and simpler reading experience.

Augmented Reality Magnification and Enhancement. Building on our work in helping blind users to read, we also applied HandSight's finger-worn camera to assist low vision users with the addition of a visual augmented reality display. In particular, we investigated the assistive potential of 3D virtual content registered in the physical environment, which previous vision enhancement systems (*e.g.*, ForeSee [235], eSight [237,238]) had not yet explored. We developed proof-of-concept AR designs that we evaluated and refined through design sessions with low vision users. Our findings were mixed; some participants were unable to use our prototype to read due to the nature of their visual impairment, while others appreciated the improved portability, privacy, and ready availability compared to their existing aids. Participants also identified advantages to our 3D AR approach compared to handheld magnification tools, including a more natural reading experience and the ability to more easily multitask,

but also some disadvantages such as a steeper learning curve and limitations of the particular hardware we used. Based on our observations and participants' feedback and open-ended ideas during design sessions, we proposed recommendations for the design of future AR vision enhancement aids.

Recognizing Location-Specific On-Body Gestures. Next, we extended our finger-worn sensing platform with additional optical and inertial sensors and implemented algorithms to recognize and localize touch gestures that users perform on their own body. On-body input offers efficient, accurate, and always-available control of mobile devices to access digital information [64,154] and could be beneficial both for visually impaired users and for eyes-free input by sighted users (*e.g.*, [40,119,131,218]). Offline evaluations demonstrated the feasibility of localizing images of small skin patches from a finger-mounted camera, and we built upon these findings to construct and evaluate a realtime on-body input system.

Findings from a user study with visually impaired participants highlighted tradeoffs in robustness and usability between different types of on-body input. Fine-grained input is efficient and discrete but challenging to classify reliably, while coarse-grained locations are easier to recognize but may also be less efficient for a new user and potentially socially unacceptable. Location-specific gestures have the potential to support efficient interaction for expert users, flexible input locations depending on user preference or situation, task-based interactions tied to intuitive locations, and relatively fine-grained input for body areas that have distinctive visual features (*e.g.*, fingertips and palm). Based on these findings, we discussed implications for the design of on-

body interfaces both in terms of usability and in which locations and gestures can be most reliably recognized across users.

Recognizing Clothing Colors and Patterns. Lastly, we applied HandSight to identifying clothing colors and visual patterns. We contributed two novel fabric texture datasets, one collected systematically using a finger-mounted camera and the other assembled from online sources. We applied transfer learning to adapt and fine-tune state-of-the-art image classifiers, demonstrating both that users could potentially train a highly accurate personalized fabric pattern classifier for the items in their own closet and that a robust generalized classifier could help to describe unfamiliar patterns while shopping. Errors were mostly attributable to the camera's proximity to the fabric, suggesting that for robust identification of patterns the camera should be positioned farther away from the surface to capture more contextual information. We implemented an interactive prototype that identifies fabric patterns and dominant colors (*e.g.*, "striped blue and white").

While preliminary, this work demonstrates feasibility and highlights the flexibility of a finger-based wearable device. Positioning the camera on the user's finger helps mitigate issues with inconsistent lighting and distance that can impact the accuracy of existing color and texture recognizers (*e.g.,* [59,227,242]) and allows for touch-based interactions with an article of clothing to better understand its appearance. Our approach should allow users to quickly explore a surface and combine their sense of touch with visual texture and color information to make informed decisions about what to wear or buy.

9.2 Limitations and Future Research Directions

This section discusses the limitations of this dissertation, both to better frame and scope our contributions and to highlight opportunities for future research. We discuss our finger-worn camera approach and potential alternatives, open questions relating to spatial exploration, more application areas for touch-based information access, alternative feedback methods, and possible extensions to other user populations.

9.2.1 Alternative or Supplementary Camera Locations

Our research focused solely on recognizing touched content and touch-based gestures using a finger-worn camera and other co-located sensors, aside from a limited qualitative comparison with a handheld smartphone camera in Chapters 4 and 5. However, as discussed in Section 9.1.1, finger-worn sensors present tradeoffs in accuracy and usability compared to other sensing approaches.

Future work should explore alternative camera locations, which may mitigate some of the limitations encountered in our research, such as insufficient contextual information and restrictive physical design options, while still supporting the positive aspects of touch-based interactions with the physical world, such as the ability to combine feedback from the system with tactile sensory information to better understand surface appearance and spatial layouts. For example, the camera could instead be positioned on the user's wrist (*e.g.*, integrated into a smartwatch, as in [195]) to reduce interference with finger movements, capture additional context around the user's touch

location, and more readily support larger and more capable components, while remaining near the interaction space for simplified sensing and flexibility.

Alternatively, combining a finger-mounted approach with a secondary camera worn on the head or upper body could balance the advantages and disadvantages of each, providing a close-up view for flexibility and robust identification of touched content and a wider view for additional contextual information and global localization and motion tracking. Upper-body cameras can use larger and higher-quality sensors (*e.g.*, depth camera, optical zoom) and could be integrated with visual or audio output (*e.g.*, built into a pair of glasses).

Another option would be to use finger-worn cameras on multiple fingers, potentially providing a wider field of view to capture additional content, supporting multitouch gestures, and enabling localized haptic feedback to enhance users' spatial exploration capabilities. Reliably integrating the information from multiple disparate sensors would likely present additional technical and reliability challenges, however, and such a system would need to be carefully designed for usability and robustness. Future work should explore the feasibility and usability of these alternatives in greater detail, and systematically compare the advantages and disadvantages of each.

9.2.2 Spatial Exploration of Documents and Other Surfaces

We conducted a preliminary investigation into touch-based exploration of document layouts (*e.g.*, the locations of paragraphs and images) in Chapter 4. Participants in our user studies successfully used HandSight to identify the locations of images in sample

magazine-style documents, trace document margins, and locate the start of paragraphs to read. Some participants found the audio cues intuitive and were able to easily locate the boundaries between columns and paragraphs, but others struggled. The ability to interpret the spatial layout of blocks of text and read sequentially line-by-line is not necessary for every document, and for simple documents it may distract from the content of the text. However, for documents with more complex layouts, such as newspapers, menus, and tables, details about the relative positions of text, images, and other document elements often contains information that is crucial for understanding the content—for example, captions beneath an image or categories and prices next to the items on a menu. Existing reading aids (*e.g.*, KNFB Reader [98]) do not provide this information, and automated methods to interpret and accurately convey complex content in an appropriate reading order are complicated and frequently inaccurate even for digital content [14,111].

Future work should investigate what types of information to convey when exploring a document. For example, beyond the simple identification of text and images that we explored, users could benefit from additional information about the purpose of a document element (*e.g.*, heading, body, caption, list) and a brief summary of its contents (*e.g.*, paragraph synopsis, image description). Future work should also investigate how best to convey spatial information, combining the user's own tactile awareness with haptic, audio, and speech cues to help users better understand a document's content.

9.2.3 Additional Applications

Our research explored three specific application areas: reading, controlling mobile devices, and identifying colors and patterns; however, HandSight has numerous additional potential applications such as exploring and interpreting inherently spatial information (*e.g.*, maps, graphs, or tables) and extending common digital interactions (*e.g.*, copy and paste, annotate, search) into the physical world. And as discussed in Section 9.2.2, additional spatial cues and information about high-level content could help when exploring printed documents.

Other researchers have begun to explore these ideas for accessing digital information using touchscreens. For example, Guidice *et al.* [52] evaluated a touchscreen vibro-audio interface to help users explore and identify non-visual information such as a bar graph, letters, or geometric shapes. User studies with their mixed-modal interface showed promising results for providing access to dynamic visual information and supporting accurate spatial learning and the development of mental representations of graphical material. A finger-worn system like HandSight with co-located sensing and feedback would be ideally suited to extend this interface into the physical world.

Similarly, several researchers (*e.g.*, [22,57,107,166,208]) have explored ways to improve the accessibility of 2D maps or art in museums by developing 3D tactile representations, refreshable tactile displays, and interactive audiovisual displays. However, these methods generally require expensive dedicated hardware or single-

purpose 3D models, limiting portability and scalability. Future work should extend our research to explore the usability and utility of touch-based access to this information.

9.2.4 Alternative Feedback Methods

We only explored limited feedback options, primarily conveying information via synthesized speech except for the simple audio and haptic cues used to guide the user's finger or identify text and images when exploring a document (Chapters 3 and 4). Alternative feedback methods could convey surface content more efficiently or intuitively. Our research covered basic auditory and haptic vibration cues to convey non-tactile information to users, but many other options exist in the fields of sonification or haptics. As discussed in Section 9.2.3, researchers have explored vibro-audio interfaces to convey graphical information on a touchscreen [52,210]. Others have explored the use of sonification and tactile displays to enable blind users to access digital map data [161,166]. To intuitively convey information about lines and shapes, future research should explore ways to seamlessly augment users' existing sense of touch with additional tactile cues. For example, researchers have explored finger-worn tactile displays to convey braille characters or other shape information [101,214]. These displays are not yet viable for end-users due to their slow response speed, expense, and power requirements, and they by necessity block the user's existing sense of touch. However, future research should explore ways to apply these and similar techniques to convey visual or digital information as tactile.

9.2.5 Extension to Other User Populations

This dissertation focused solely on visually impaired users, as they benefit most from touch-based access to non-tactile information. However, our research could also extend to other populations as well. First and most obviously, users with color vision deficiencies could benefit from a robust interactive color identifier, which our approach could readily support. While numerous smartphone color identification apps are available to assist with distant color identification, a wearable touch-based system would for example allow for quick spot checks when shopping, cooking, or attempting to interpret the colors used in images and graphs. Second, touch-based reading could be helpful as a teaching aid to support readers in early education, with dyslexia, or other users who cannot read unassisted. HandSight would directly allow users to associate spoken words with their visual appearance as they move their fingers across a page and could readily support users' existing materials. Third, our work to support on-body interactions could also offer eyes-free input for any user, visually impaired or sighted. On-body gestures could provide an alternative to existing touchscreen and voice controls for efficient, accurate, and always available control of computers and mobile devices. Similarly to Magic Finger [228], our approach could potentially support touch-based interactions on any surface. And fourth, touch-based interactions augmented with haptic and audio feedback could be useful for augmented and virtual reality as an input and feedback method in place of a controller. For example, a finger-worn camera could be used to identify real-world objects that the user is touching, providing tangible interactions with virtual augmentations.

9.3 Final Remarks

This dissertation provides insights into issues relating to the design and implementation of a wearable system to support touch-based access to information. We constructed HandSight—which augments visually impaired users' fingers with sensing and feedback capabilities—and explored its potential through three specific application areas: reading and exploring printed materials, controlling mobile devices to access digital information, and identifying clothing colors and patterns. Our research is an early exploration into finger-worn assistive cameras and many open questions and areas for future work remain; however, we have demonstrated the potential advantages of systems like HandSight, that include simplified sensing and processing, flexible input location and intuitive camera aiming, and integrated knowledge from the system's feedback and the user's own sense of touch. We believe that this dissertation achieves our goal of increasing the accessibility of information for people with visual impairments and that it serves as a first step toward a general system for supporting touch-based interactions and non-tactile information access on any surface—benefitting both visually impaired and sighted users.

Appendix A

In this appendix we list the text of the documents used in Chapter 4 for Reading Studies I and II, along with associated comprehension questions. We adapted six test documents from a Grade 8 Iowa Test of Basic Skills practice book [167]. The original text was modified slightly for length and to ensure clarity with our speech synthesis engine (*e.g.,* removing unnecessary proper nouns). We created three additional training documents of a similar length and reading level, as well as a two-column magazine document for testing KNFB Reader iOS, using documents adapted from articles in USA Today.

Study I, Training Document (plain, both conditions):

Scientists counting Antarctica's emperor penguins from space have found twice as many of them as expected. Using high-resolution satellite images to study each of 44 colonies around the coastline of Antarctica, experts said Friday that they put the total emperor penguin population at 595 thousand, or roughly double previous estimates of 270 thousand to 350 thousand. Seven of the colonies had never been seen before.

Satellite technology was a boon for researchers; visiting dozens of remote colonies in temperatures as low as minus 58 degrees is expensive, dangerous and time-consuming. With their distinctive black and white plumage, emperor penguins stand out against the snow, making them clearly visible on satellite images.

Study I, Test Document 1 (plain, first condition):

People have used coins as a means of exchange for thousands of years. Valued for their craftsmanship and purchasing power, coins have been collected in great numbers throughout history and buried for safekeeping. Because stores of coins gathered and hidden in this manner lie untouched for many years, they can reveal a great deal about a given culture.

Coins are useful in revealing many aspects of a culture. They can provide clues about when a given civilization was wealthy and when it was experiencing a depression. Wealthy nations tend to produce a greater number of coins made from richer materials. The distribution of coins can also reflect the boundaries of an empire and the trade relationships within it. Roman imperial gold coins found in India, for example, indicate the Romans purchased goods from the East.

The way the coins themselves are decorated sometimes provides key information about a culture. Many coins are stamped with a wealth of useful historical evidence, including portraits of political leaders, important buildings and sculptures, mythological and religious figures, and useful dates. Some coins, such as many from ancient Greece, can be considered works of art themselves and reflect the artistic achievement of the civilization as a whole.

Information gathered from old coins by historians is most useful when placed alongside other historical documents, such as written accounts or data from archeological digs. Combined with these other pieces of information, coins can help historians reconstruct the details of lost civilizations.

Comprehension Questions:

1. Which of the following do coins reveal about a civilization?
    a. The average cost of clothing
    b. Information about its economy
    c. Its farming techniques
2. What is the main idea of the passage?
    a. How difficult it is to find old coins
    b. How coins reflect the artistic achievements of a culture
    c. How coins can tell us about ancient civilizations

Study I, Test Document 2 (magazine, first condition):

Despite the stubborn, widespread opinion that animals don't feel emotions in the same way that humans do, many animals have been observed to demonstrate a capacity for joy. People have often seen animals evincing behavior that can only be taken to mean they are pleased with what life has brought them in that particular moment.

A chimpanzee named Nim was raised by a human family for the first year and a half of his life. After that time, Nim was separated from them for two and a half years. On the day that Nim was reunited with his human family, he smiled, shrieked, pounded the ground, and looked from one member of the family to the next. Still smiling and shrieking, Nim went around hugging each member of the family. He played with and groomed each member of the family for almost an hour before the family had to leave. People who were familiar with Nim's behavior said they had never seen him smile for such a long period of time.

Comprehension Questions:

1. What is the purpose of the second paragraph?
    a. To criticize Nim's human family for abandoning him
    b. To show how well Nim's human family treated him
    c. To demonstrate that animals have the ability to feel joy
2. Why did Nim shriek and pound the ground?
    a. He was overjoyed to see the family again.
    b. He was hungry and wanted to be fed.
    c. He was angry with the family for leaving him.

Study I, Test Document 3 (plain, second condition):

Born in Spanish Harlem in the late 1950s, Raphael Sanchez learned at an early age to listen to the many voices of the city. It was as a boy in Harlem that he developed the powers of observation that would later make his

writing truly great. In the 1970s, Raphael went to Columbia University, where he was exposed to a literary tradition. While his university education gave his writing new depth, the raw energy of the streets has always served as the primary fuel for his writing. This is what gives his works passion and power.

Raphael once told me that in order to escape from life he turns to books, and in order to escape from books he turns to life. It is this balance of the sights, sounds, and smells of the street with the perspective gained from his formal education that has made Raphael popular with both critics and regular readers alike.

For those of us who have read and admired his work, it seems natural that Raphael has won so many awards. He deserves them, and his humility in accepting them has been refreshing. When he received the Writer's Quill Award two weeks ago, for example, he told the audience, "This award is not really mine. It belongs to all the million things that have inspired me.

"That is the kind of man I am introducing to you this evening. He is a man who has been inspired by a million things, and he is a man who has provided inspiration to a million people. Ladies and gentlemen, it is my great honor to present to you, Raphael Sanchez.

Comprehension Questions:

1. Which of these best describes why Raphael Sanchez's writing is so popular with critics and regular readers?

    a. It has won the Writer's Quill Award.

    b. It reflects both scholarship and city experience.

    c. It is based on his experiences at Columbia University.

2. What does Raphael Sanchez mean when he says, "This award is not really mine"?

    a. He owes everything to the people and things that inspired him.

    b. He does not believe in the value of awards.

    c. He feels Columbia University should be given an award too.

Study I, Test Document 4 (magazine, second condition):

In the 1800s, most geologists thought the sea floor was a lifeless expanse of mud, sediment, and the decaying remains of dead organisms. They thought that, with the exception of some volcanic islands, the bottom of the sea had no major geographic features, such as peaks or valleys.

In the mid-nineteenth century, ships depth-sounding the ocean floor with sonar for a transatlantic telegraph cable made some interesting discoveries. To geologists' surprise, the ocean floor was found to be made up of long mountain ranges and deep valleys and troughs. Another surprise finding in the Atlantic was the existence of basalt, a volcanic rock thought only to exist in the Pacific Ocean. The presence of basalt in the

Atlantic was a clue that volcanic activity occurs at the bottom of the sea. This and other discoveries, many of them accidental in the beginning, were signals to geologists that their knowledge of the sea floor was very limited.

Comprehension Questions:

1. The discovery of basalt in the Atlantic Ocean suggested that

    a. Iron, zinc, and gold would be found nearby.

    b. Geologists still had much to learn about the ocean floor.

    c. The Atlantic was deeper than previously believed.

2. How did ships in the mid-nineteenth century measure the ocean's depth?

    a. By sending down scuba divers

    b. By bouncing sound waves off the sea bottom

    c. By photographing the sea floor with special cameras

Study II, Training Document 1 (plain, HandSight):

When Mary Smallenburg opened a package from her mother to find cereal and ramen noodles, she burst into tears. Without it, she wouldn't be able to feed her four children. It got to the point where I opened my pantry and there was nothing. Nothing. What was I going to feed my kids? Smallenburg says, adjusting a bag of fresh groceries on her arm.

Smallenburg's family is one of 50 military families that regularly visit the Lorton Community Action Center food bank. Volunteers wave a familiar hello as she walks in the door. None of what we have been through has been expected, Smallenburg says. Three of her four children have special needs and her husband is deployed in Korea. The last few months actually, coming here has been a godsend.

Nationwide, 25 percent of military families, 620,000 households, need help putting food on the table, according to a study by Feeding America, a network of 200 food banks. The results are alarming, says Bob Aiken, chief executive officer of Feeding America. It means that people in America have to make trade-offs. They have to pick between buying food for their children or paying for utilities, rent and medicine.

One in seven Americans, 46 million people, rely on food pantries and meal service programs to feed themselves and their families, the study found.

Study II, Test Document 1 (plain, HandSight):

Henry Ford and his Model T automobile changed the face of America. His horseless carriage contributed to a movement from rural to urban and to the development of an industrial economy.

In 1903, Ford Motor Company was officially formed, and in 1908, Ford announced the birth of the Model T. He told the nation, I will build a car

for the great multitude. This was a bold announcement, since most manufacturers planned to build only luxury cars for the very wealthy.

His idea worked. Ford's Model T was a hit with the American public, and demand grew with each passing year. In the course of nineteen years, around fifteen million Model T cars were sold in the United States, nearly one million in Canada, and another 250,000 in Great Britain. All told, these numbers equaled half the total number of automobiles manufactured in the world at that time.

The methods of production Ford used were revolutionary. Ford's assembly line could churn out the frame of a Model T in about six hours. This high rate of speed was made possible by the division of labor. Instead of one person controlling production from start to finish, the labor was divided into smaller and smaller tasks, with each person performing the same task all day long.

By 1927 the era of the Model T was coming to a close. General Motors, a major competitor, was producing better cars for only slightly more money. Customers with an eye for new styles just didn't see the appeal of the plain Model T.

Comprehension Questions:

1. According to the passage, why was the Model T more popular than other cars available at the same time?

a. It looked like a buggy.

b. It was more spacious.

c. It was less expensive.

d. It was more stylish.

2. Which of the following best describes Ford according to the article?

a. A poor businessman

b. A visionary

c. A follower

d. A great metal worker

3. What led to the downfall of the Model T?

a. It was not very well made.

b. Its price went up.

c. Other competition emerged.

d. Many of Ford's workers quit their jobs.

Study II, Training Document 2 (plain, KNFB Reader iOS):

Here's a tip. Don't stress over tipping.

Restaurant tips are more modest in Europe than in America. In most places, 10 percent is a big tip. If your bucks talk at home, muzzle them on your travels. As a matter of principle, if not economy, the local price

should prevail. Please believe me, tipping 15 percent or 20 percent in Europe is unnecessary, if not culturally ignorant.

Virtually anywhere in Europe, you can do as the Europeans do and, if you're pleased with the service, add a euro or two for each person in your party. In very touristy areas, some servers have noticed the American obsession with overtipping, and might hope for a Yankee-size tip. But the good news is that European servers and diners are far more laid-back about all this than we are. The stakes are low, and it's no big deal if you choose the wrong amount. And note that tipping is an issue only at restaurants that have waiters and waitresses. If you order your food at a counter, don't tip.

At table-service restaurants, the tipping etiquette and procedure vary slightly from country to country. But in general, European servers are well paid, and tips are considered a small bonus, to reward great service or for simplicity in rounding the total bill to a convenient number. In many countries, 5 percent to 10 percent is sufficient.

Study II, Test Document 2 (plain, KNFB Reader iOS):

A clone is a life form engineered in a lab environment to be identical to another, through a process of asexual, or nonsexual, reproduction. This process of creating a new life form, called genetic engineering, can be useful in creating individuals of a given species that represent the best

possible genetic traits of that species. People who work with plants have long used cloning techniques to create better strains of trees, fruits, and vegetables. The Macintosh apple, for example, was created by cloning techniques, and it supposedly represents the best qualities of all apple types.

In July of 1996, a group of Scottish scientists made a breakthrough by successfully cloning a sheep from the cells of another adult sheep. After scraping cells from the udder of one sheep, the scientists introduced the nucleus of one of these cells into the unfertilized eggs of a different sheep. Then, they placed the egg, which had begun to divide, into the uterus of a third sheep. The result was Dolly, a healthy sheep who was born in the natural way from the third sheep. Dolly was almost identical to the sheep from whose skin cells she had been formed.

In 1997 Dolly's story was widely publicized in the media, and her existence resparked a continuing debate about the use of cloning techniques on humans. Some people claim that genetic engineering should not be used on humans under any circumstances. Others urge slowness. They insist that if genetic engineering is to be used, there are many questions that need to be answered first.

Comprehension Questions:

1. What event revived the debate about cloning?

a. The cloning of plants

b. The cloning of a sheep

c. The cloning of amphibians

d. The future plan to clone human organs

2. According to the passage, how many sheep did it take to produce Dolly?

a. One

b. Two

c. Three

d. Four

3. In paragraph 1, the author mentions the Macintosh apple as an example of

a. a case when cloning produced an improved organism.

b. a case when cloning failed.

c. a case when cloning created a controversy.

d. a case when cloning went too far.

Study II, Test Document 3 (magazine, KNFB Reader iOS):

Let them sleep! That's the message from the nation's largest pediatrician group, which, in a new policy statement, says delaying the start of high school and middle school classes to 8:30 a.m. or later is an effective

countermeasure to chronic sleep loss and the epidemic of delayed, insufficient, and erratic sleep patterns among the nation's teens.

Multiple factors, including biological changes in sleep associated with puberty, lifestyle choices, and academic demands, negatively impact teens' ability to get enough sleep, and pushing back school start times is key to helping them achieve optimal levels of sleep, 8 and a half to 9 and a half hours a night, says the American Academy of Pediatrics statement, released Monday and published online in Pediatrics.

Just 1 in 5 adolescents get nine hours of sleep on school nights, and 45 percent sleep less than eight hours, according to a 2006 poll by the National Sleep Foundation (NSF).

As adolescents go up in grade, they're less likely with each passing year to get anything resembling sufficient sleep, says Judith Owens, director of sleep medicine at Children's National Medical Center in Washington, D.C., and lead author of the AAP statement. By the time they're high school seniors, the NSF data showed they were getting less than seven hours of sleep on average.

Chronic sleep loss in children and adolescents can, without hyperbole, really be called a public health crisis, Owens says.

Appendix B

In this appendix we list the text of the subjective questionnaires administered in Reading Studies I and II (Chapter 4). For ease of use questions, the choices were (1) Very difficult, (2) Somewhat difficult, (3) Neutral, (4) Somewhat easy, or (5) Very easy. For comparison questions, the choices were (1) Much worse, (2) Somewhat worse, (3) About the same, (4) Somewhat better, or (5) Much better.

Study I, after each directional guidance condition:

1. How easy or difficult was it to follow a line of text with your finger? Why?
2. How easy or difficult was it to read and understand the text given this feedback? Why?
3. Do you feel like the feedback direction was correct, or did if feel backwards to you?
4. Do you have any other comments about what was good or bad about this type of feedback?

Study I, end of study:

1. Overall, how easy or difficult was it to find the beginning of the text?
2. How easy or difficult was it to find the beginning of each line?
3. How easy or difficult was it to notice the end of a line?
4. How easy or difficult was it to notice the end of a paragraph?
5. How easy or difficult was it to find the beginning of the next column?

6. You have tried two different types of feedback. Which did you prefer more? Why?
7. Overall, how was your experience of our system compared to how you would normally read braille? Why?
8. Overall, how was your experience of our system compared to how you would normally use a screen reader? Why?
9. Overall, how was your experience of our system compared to how you would normally read printed documents? Why?
10. Do you have any questions, suggestions for improvement, or other comments?

Study II, after HandSight tasks:

1. How easy or difficult was it to find the beginning of the text?
2. How easy or difficult was it to find the beginning of each line?
3. How easy or difficult was it to notice the end of a line?
4. How easy or difficult was it to notice the end of a paragraph?
5. How easy or difficult was it to follow a line of text with your finger?
6. How easy or difficult was it to understand the feedback?
7. Overall, how easy or difficult was it to read and understand the text given this feedback?
8. Overall, how was your experience with the app compared to how you would normally read braille?

9. Overall, how was your experience with the app compared to how you would normally read printed documents?

10. Do you have any other comments about what was good or bad about this type of feedback?

Study II, after KNFB Reader iOS tasks:

1. Overall, how easy or difficult was it to read and understand the text using the app?

2. What, if anything, did you like about using the app?

3. What, if anything, did you dislike about using the app?

4. Overall, how was your experience with the app compared to how you would normally read braille documents?

5. Overall, how was your experience with the app compared to how you would normally read printed documents?

6. Overall, how was your experience with the app compared to reading with HandSight?

7. Do you have any questions, suggestions for improvement, or other comments about the KNFB Reader app?

Appendix C

This appendix includes additional information about the features used for TouchCam Offline in Chapter 7, and a table summarizing statistically significant comparisons between combinations of sensors.

Localization Features		Motion Features	
IR	2 raw IR sensor readings	IR	70 features: 50 resampled points + 5 summary statistics × 4 windows
IMUs	4D orientation vector (quaternion) for each IMU	IMUs	639 features: 3 sensors × [3 axes × (50 resampled points + 5 summary statistics × 4 windows) + 3 correlation values]
Camera	LBP texture histogram with 1792 bins (14 patterns x 16 variances x 8 scales) 2D Gabor keypoints, variable number per image	Camera	140 features: 2 axes × (50 resampled points + 5 summary statistics × 4 windows)

Table C.1: Summary of localization and motion features extracted from each sensor for TouchCam Offline (Chapter 7, On-body Input Study I).

Single Sensors	t_{23}	p	d
IR vs. W	-17.90	<0.001	-3.65
IR vs. F	-17.16	<0.001	-3.50
IR vs. C	-10.38	<0.001	-2.12
W vs. C	4.34	<0.001	0.89
F vs. C	3.63	0.005	0.74

Two Sensors	t_{23}	p	d
IR+C vs. F+W	-8.85	<0.001	-1.81
IR+C vs. W+C	-7.46	<0.001	-1.52
IR+W vs. IR+C	7.39	<0.001	1.51
IR+F vs. IR+C	6.72	<0.001	1.37
IR+C vs. F+C	-6.59	<0.001	-1.35
F+W vs. F+C	6.20	<0.001	1.27
IR+F vs. F+W	-4.67	<0.001	-0.95
F+W vs. W+C	3.73	0.008	0.76
IR+W vs. F+W	-3.23	0.036	-0.66

Best Single (W) vs. Two Sensors	t_{23}	p	d
W vs. IR+C	5.73	<0.001	1.17
W vs. W+C	-4.81	<0.001	-0.98
W vs. IR+W	-4.81	<0.001	-0.98
W vs. F+W	-4.75	<0.001	-0.97

Three Sensors	t_{23}	p	d
IR+F+W vs. IR+F+C	6.46	<0.001	1.32
IR+F+C vs. F+W+C	-4.71	<0.001	-0.96
IR+F+W vs. IR+W+C	3.64	0.003	0.74
IR+F+W vs. F+W+C	-3.17	0.016	-0.65

Best Three (IR+F+W) vs. All Four Sensors	t_{23}	p	d
IR+F+W vs. All	-2.13	0.044	-0.44

Best Two (F+W) vs. Three Sensors	t_{23}	p	d
F+W vs. IR+F+C	4.17	<0.001	0.85
F+W vs. IR+F+W	-2.99	0.020	-0.61
F+W vs. F+W+C	-2.67	0.027	-0.55
F+W vs. IR+W+C	-2.67	0.026	0.49

Table C.2: Statistically significant comparisons between combinations of sensors used in On-body Input Study I (Chapter 7).

Bibliography

[1] Ibrahim SI Abuhaiba. 2006. Efficient OCR using Simple Features and Decision Trees with Backtracking. *Arabian Journal for Science and Engineering* 31, 2: 223–244.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and Sabine Su?sstrunk. 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 11: 2274–2282. https://doi.org/10.1109/TPAMI.2012.120

[3] Dustin Adams, Lourdes Morales, and Sri Kurniawan. 2013. A qualitative study to support a blind photography mobile application. *Proc. PETRA 2013*: 1–8. https://doi.org/10.1145/2504335.2504360

[4] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. 2004. Face Recognition with Local Binary Patterns. In *Computer Vision - ECCV 2004*. 469–481. https://doi.org/10.1007/978-3-540-24670-1_36

[5] Daniel Ashbrook, Patrick Baudisch, and Sean White. 2011. Nenya: Subtle and Eyes-free Mobile Input with a Magnetically-tracked Finger Ring. In *Proc. CHI 2011*, 2043–2046. https://doi.org/10.1145/1978942.1979238

[6] Douglas Astler, Hayato Unno, Carol Wang, Khadija Zaidi, Xuemin Zhang, Cha-Min Tang, Harrison Chau, Kailin Hsu, Alvin Hua, Andrew Kannan, Lydia Lei, Melissa Nathanson, Esmaeel Paryavi, and Michelle Rosen. 2011. Increased accessibility to nonverbal communication through facial and expression recognition technologies for blind/visually impaired subjects. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '11*, 259. https://doi.org/10.1145/2049536.2049596

[7] AWAIBA. NanEye Medical Image Sensors. Retrieved from http://www.awaiba.com/en/products/medical-image-sensors/

[8] Shiri Azenkot and Nicole B Lee. 2013. Exploring the use of speech input by blind people on mobile devices. In *Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility*, Article No. 11.

[9] Shiri Azenkot, Kyle Rector, Richard E. Ladner, and Jacob O. Wobbrock. 2012. PassChords: secure multi-touch authentication for blind people. In *Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility*, 159–166.

[10] Ronald T. Azuma. 1997. A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments* 6, 4: 355–385. https://doi.org/10.1162/pres.1997.6.4.355

[11] Jeffrey P. Bigham, Samual White, Tom Yeh, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, and Brandyn White. 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*, 333–342. https://doi.org/10.1145/1866029.1866080

[12] J C Bliss. 1969. A relatively high-resolution reading aid for the blind. *Man-Machine Systems, IEEE Transactions on* 10, 1: 1–9. https://doi.org/10.1109/TMMS.1969.299874

[13] Matthew N Bonner, Jeremy T Brudvik, Gregory D Abowd, and W. Keith Edwards. 2010. No-Look Notes: accessible eyes-free multi-touch text entry. In *Proceedings of Pervasive Computing*, 409–426. Retrieved November 10, 2010 from http://www.springerlink.com/index/3R67570211465164.pdf

[14] Yevgen Borodin, Jeffrey P. Bigham, Glenn Dausch, and I. V. Ramakrishnan. 2010. More than meets the eye: a survey of screen-reader browsing strategies. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A) - W4A '10*, 1. https://doi.org/10.1145/1805986.1806005

[15] Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. 2013. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, 1225. https://doi.org/10.1145/2441776.2441915

[16] 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*, 2117–2126. Retrieved November 18, 2014 from http://dl.acm.org.proxy-um.researchport.umd.edu/citation.cfm?id=2470654.2481291

[17] Brytech. Brytech Color Teller. Retrieved from http://www.brytech.com/colorteller/

[18] David S. Burch and Dianne T.V. Pawluk. 2009. A cheap, portable haptic device for a method to relay 2-D texture-enriched graphical information to individuals who are visually impaired. In *Proceeding of the eleventh*

international ACM SIGACCESS conference on Computers and accessibility - ASSETS '09, 215. https://doi.org/10.1145/1639642.1639682

[19] Michele A Burton. 2011. Fashion for the Blind: A Study of Perspectives. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '11), 315–316. https://doi.org/10.1145/2049536.2049625

[20] Michele A Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P Bigham, and Amy Hurst. 2012. Crowdsourcing Subjective Fashion Advice Using VizWiz: Challenges and Opportunities. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '12), 135–142. https://doi.org/10.1145/2384916.2384941

[21] John Canny. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8, 6: 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

[22] Virginio Cantoni, Luca Lombardi, Alessandra Setti, Stanislav Gyoshev, Dimitar Karastoyanov, and Nikolay Stoimenov. 2018. Art Masterpieces Accessibility for Blind and Visually Impaired People. . Springer, Cham, 267–274. https://doi.org/10.1007/978-3-319-94274-2_37

[23] Steve Caperna, Christopher Cheng, Junghee Cho, Victoria Fan, Avishkar Luthra, Brendan O'Leary, Jansen Sheng, Andrew Sun, Lee Stearns, Roni Tessler, Paul Wong, and Jimmy Yeh. 2009. A navigation and object location device for the blind. University of Maryland, College Park.

[24] M Capp and P Picton. 2000. The optophone: an electronic blind aid. *Engineering Science and Education Journal* 9, 3: 137–143. https://doi.org/10.1049/esej:20000306

[25] Liwei Chan, Yi-Ling Chen, Chi-Hao Hsieh, Rong-Hao Liang, and Bing-Yu Chen. 2015. CyclopsRing: Enabling Whole-Hand and Context-Aware Interactions Through a Fisheye Ring. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology - UIST '15*, 549–556. https://doi.org/10.1145/2807442.2807450

[26] Liwei Chan, Chi-Hao Hsieh, Yi-Ling Chen, Shuo Yang, Da-Yuan Huang, Rong-Hao Liang, and Bing-Yu Chen. 2015. Cyclops: Wearable and single-piece full-body gesture input devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 3001–3009. https://doi.org/10.1145/2702123.2702464

[27] Liwei Chan, Rong-Hao Liang, Ming-Chang Tsai, Kai-Yin Cheng, Chao-Huai

Su, Mike Y Chen, Wen-Huang Cheng, and Bing-Yu Chen. 2013. FingerPad: Private and Subtle Interaction Using Fingertips. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (UIST '13), 255–260. https://doi.org/10.1145/2501988.2502016

[28] Xiangrong Chen and A L Yuille. 2004. Detecting and reading text in natural scenes. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, II-366-II-373 Vol.2. https://doi.org/10.1109/CVPR.2004.1315187

[29] Christopher Cheng, Brendan O'Leary, Lee Stearns, Steve Caperna, Junghee Cho, Victoria Fan, Avishkar Luthra, Andrew Sun, Roni Tessler, Paul Wong, Jimmy Yeh, Bobby Bobo, Rama Chellappa, and Cha-Min Tang. 2008. Developing a Real-Time Identify-and-Locate System for the Blind. In *Workshop on Computer Vision Applications for the Visually Impaired*.

[30] Michał Choraś and Rafał Kozik. 2012. Contactless palmprint and knuckle biometrics for mobile devices. *Pattern Anal. Appl.* 15, 1: 73–85. https://doi.org/10.1007/s10044-011-0248-4

[31] Yung-Long Chu, Hung-En Hsieh, Wen-Hsiung Lin, Hui-Ju Chen, and Chien-Hsing Chou. 2017. Chinese FingerReader: a wearable device to explore Chinese printed text. In *ACM SIGGRAPH 2017 Posters on - SIGGRAPH '17*, 1–2. https://doi.org/10.1145/3102163.3102206

[32] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing Textures in the Wild. In *Proceedings of CVPR 2014*, 3606–3613.

[33] Franklin S Cooper, Jane H Gaitenby, and Patrick W Nye. 1983. *Evolution of reading machines for the blind: Haskins Laboratories' research as a case history*. Haskins Laboratories.

[34] James Coughlan, James Coughlan, Roberto M, and Huiying Shen. 2006. Cell phone-based wayfinding for the visually impaired. *ST INT. WORKSHOP ON MOBILE VISION* 1. Retrieved October 30, 2016 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.87.8102

[35] Michael D. Crossland, Rui S. Silva, and Antonio F. Macedo. 2014. Smartphone, tablet computer and e-reader use by people with vision impairment. *Ophthalmic and Physiological Optics* 34, 5: 552–557. https://doi.org/10.1111/opo.12136

[36] Michael Cutter and Roberto Manduchi. 2015. Towards Mobile OCR: How To Take a Good Picture of a Document Without Sight. *Proceedings of the ACM*

Symposium on Document Engineering. ACM Symposium on Document Engineering 2015: 75–84. Retrieved January 12, 2016 from http://dl.acm.org/citation.cfm?id=2682571.2797066

[37] EE Fournier D'Albe. 1914. On a type-reading optophone. *Proceedings of the Royal Society of London. Series A* 90, 619: 373–375.

[38] Dar-Shyang Lee and S.N. Srihari. 1995. A theory of classifier combination: the neural network approach. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 42–45. https://doi.org/10.1109/ICDAR.1995.598940

[39] Mohammad Omar Derawi, Bian Yang, and Christoph Busch. 2012. Fingerprint Recognition with Embedded Cameras on Mobile Phones. In *Security and Privacy in Mobile Info. and Com. Sys.* Springer, 136–147. https://doi.org/10.1007/978-3-642-30244-2_12

[40] Niloofar Dezfuli, Mohammadreza Khalilbeigi, Jochen Huber, Florian Müller, and Max Mühlhäuser. 2012. PalmRC: Imaginary Palm-Based Remote Control for Eyes-free Television Interaction. In *Proc. EuroITV '12*, 27. Retrieved August 13, 2015 from http://dl.acm.org/citation.cfm?id=2325616.2325623

[41] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: a deep convolutional activation feature for generic visual recognition. *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, I-647. Retrieved June 13, 2018 from https://dl.acm.org/citation.cfm?id=3044879

[42] Julien Doublet, Marinette Revenu, and Olivier Lepetit. 2007. Robust GrayScale Distribution Estimation for Contactless Palmprint Recognition. In *IEEE Conference on Biometrics: Theory, Applications, and Systems 2007*, 1–6. https://doi.org/10.1109/BTAS.2007.4401935

[43] Sarah L Dowhower. 1989. Repeated reading: research into practice. *The Reading Teacher* 42, 7: 502–507. Retrieved from http://www.jstor.org/stable/20200198

[44] Murat Ekinci and Murat Aykut. 2008. Palmprint Recognition by Applying Wavelet-Based Kernel PCA. *Computer Science and Technology* 23, 107: 851–861.

[45] Enhanced Vision Systems. JORDY: Joint Optical Reflective Display. Retrieved from https://www.enhancedvision.com/low-vision-product-line/jordy.html

[46] Eryun Liu, A. K. Jain, and Jie Tian. 2013. A Coarse to Fine Minutiae-Based Latent Palmprint Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 10: 2307–2322. https://doi.org/10.1109/TPAMI.2013.39

[47] N Ezaki, K Kiyota, B T Minh, M Bulacu, and L Schomaker. 2005. Improved text-detection methods for a camera-based text reading system for blind persons. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, 257–261 Vol. 1. https://doi.org/10.1109/ICDAR.2005.137

[48] Martin A. Fischler and Robert C. Bolles. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM* 24, 6: 381–395. https://doi.org/10.1145/358669.358692

[49] Yoav Freund and Robert E. Schapire. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT 1995*. Springer, Berlin, Heidelberg, 23–37. https://doi.org/10.1007/3-540-59119-2_166

[50] Brian Frey, Caleb Southern, and Mario Romero. 2011. BrailleTouch: mobile texting for the visually impaired. In *Proceedings HCI International*, 19–25.

[51] Vincent Gaudissart, Silvio Ferreira, Céline Thillou, and Bernard Gosselin. 2004. SYPOLE: mobile reading assistant for blind people. In *9th Conference Speech and Computer (SPECOM)*. Retrieved from http://www.isca-speech.org/archive_open/specom_04/spc4_538.pdf

[52] Nicholas A. Giudice, Hari Prasath Palani, Eric Brenner, and Kevin M. Kramer. 2012. Learning non-visual graphical information using a touch-based vibro-audio interface. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '12*, 103. Retrieved August 13, 2015 from http://dl.acm.org/citation.cfm?id=2384916.2384935

[53] Mayank Goel, Shwetak N. Patel, Eric Whitmire, Alex Mariakakis, T. Scott Saponas, Neel Joshi, Dan Morris, Brian Guenter, Marcel Gavriliu, and Gaetano Borriello. 2015. HyperCam: Hyperspectral Imaging for Ubiquitous Computing Applications. In *Proceedings of UbiComp '15*, 145–156. Retrieved October 20, 2015 from http://dl.acm.org/citation.cfm?id=2750858.2804282

[54] Louis H. Goldish and Harry E. Taylor. 1973. The Optacon: a valuable device for blind persons. *New Outlook for the Blind*. Retrieved November 23, 2013 from http://eric.ed.gov/?id=EJ096181

[55] Daniel Goldreich and Ingrid M. Kanics. 2003. Tactile Acuity is Enhanced in Blindness. *J. Neurosci.* 23, 8: 3439–3445. Retrieved November 24, 2013 from http://www.jneurosci.org/content/23/8/3439.abstract

[56] Google. Google Glass. Retrieved from https://www.google.com/glass/start/

[57] Timo Götzelmann. 2016. LucentMaps: 3D Printed Audiovisual Tactile Maps for Blind and Visually Impaired People. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '16*, 81–90. https://doi.org/10.1145/2982142.2982163

[58] Joan M. Greenbaum and Morten Kyng. 1991. *Design at work : cooperative design of computer systems*. L. Erlbaum Associates. Retrieved April 16, 2018 from https://dl.acm.org/citation.cfm?id=125470

[59] GreenGar Studios. Color Identifier. Retrieved from https://itunes.apple.com/us/app/color-identifier/id363346987?mt=8

[60] Tiago Guerreiro, Tiago Guerreiro, Hugo Nicolau, and Joaquim A. Jorge. From tapping to touching: Making touch screens accessible to blind users. *IEEE MULTIMEDIA*: 48--50. Retrieved November 5, 2016 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.718.6436

[61] Anhong Guo, Xiang "Anthony" Chen, Haoran Qi, Samuel White, Suman Ghosh, Chieko Asakawa, and Jeffrey P. Bigham. 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 - UIST '16*, 651–664. https://doi.org/10.1145/2984511.2984518

[62] Zhenhua Guo, Lei Zhang, and David Zhang. 2010. Rotation invariant texture classification using LBP variance (LBPV) with global matching. *Pattern Recognition* 43, 3: 706–719. https://doi.org/10.1016/j.patcog.2009.08.017

[63] Sean Gustafson, Daniel Bierwirth, and Patrick Baudisch. 2010. Imaginary Interfaces: Spatial Interaction with Empty Hands and Without Visual Feedback. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology* (UIST '10), 3–12. https://doi.org/10.1145/1866029.1866033

[64] Sean G Gustafson, Bernhard Rabe, and Patrick M Baudisch. 2013. Understanding Palm-based Imaginary Interfaces: The Role of Visual and Tactile Cues when Browsing. In *Proceedings of CHI '13* (CHI '13), 889–898. https://doi.org/10.1145/2470654.2466114

[65] Sean Gustafson, Christian Holz, and Patrick Baudisch. 2011. Imaginary Phone:

Learning Imaginary Interfaces by Transferring Spatial Memory from a Familiar Device. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (UIST '11), 283–292. https://doi.org/10.1145/2047196.2047233

[66] R Harper, L Culham, and C Dickinson. 1999. Head mounted video magnification devices for low vision rehabilitation: a comparison with existing technology. *The British journal of ophthalmology* 83, 4: 495–500. https://doi.org/10.1136/BJO.83.4.495

[67] Harrison, H Benko, and a Wilson. 2011. OmniTouch: wearable multitouch interaction everywhere. *Proceedings of UIST 2011*: 441–450. https://doi.org/10.1145/2047196.2047255

[68] ChandraM. Harrison. 2004. Low-vision reading aids: reading as a pleasurable experience. *Personal and Ubiquitous Computing* 8, 3–4: 213–220. https://doi.org/10.1007/s00779-004-0280-0

[69] Chris Harrison, Desney Tan, and Dan Morris. 2010. Skinput: Appropriating the Body As an Input Surface. In *Proceedings of CHI '10* (CHI '10), 453–462. https://doi.org/10.1145/1753326.1753394

[70] Chris Harrison and Andrew D Wilson. 2011. OmniTouch: wearable multitouch interaction everywhere. In *Proc. of UIST '11*, 441–450.

[71] Richard Hartley and Andrew Zisserman. 2003. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

[72] Sharon A. Haymes, Alan W. Johnston, and Anthony D. Heyes. 2002. Relationship between vision impairment and ability to perform activities of daily living. *Ophthalmic and Physiological Optics* 22, 2: 79–91. Retrieved November 22, 2013 from http://doi.wiley.com/10.1046/j.1475-1313.2002.00016.x

[73] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

[74] A. Helal, S.E. Moore, and B. Ramachandran. 2001. Drishti: an integrated navigation system for visually impaired and disabled. In *Proceedings Fifth International Symposium on Wearable Computers*, 149–156. https://doi.org/10.1109/ISWC.2001.962119

[75] J. A. Hesch and S. I. Roumeliotis. 2010. Design and Analysis of a Portable

Indoor Localization Aid for the Visually Impaired. *The International Journal of Robotics Research* 29, 11: 1400–1415. Retrieved November 22, 2013 from http://ijr.sagepub.com/content/29/11/1400.short

[76] David W. Hislop, B. L. Zuber, and John L. Trimble. 1983. Characteristics of reading rate and manual scanning patterns of blind Optacon readers. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 25, 4: 379–389. Retrieved November 23, 2013 from http://hfs.sagepub.com/content/25/4/379.short

[77] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*: 65–70.

[78] Jonggi Hong, Lee Stearns, Jon Froehlich, David Ross, and Leah Findlater. 2016. Evaluating Angular Accuracy of Wrist-based Haptic Directional Guidance for Hand Movement. In *Proceedings of the 42Nd Graphics Interface Conference* (GI '16), 195–200. https://doi.org/10.20380/GI2016.25

[79] Samantha Horvath, John Galeotti, Bing Wu, Roberta Klatzky, Mel Siegel, and George Stetten. 2014. FingerSight: Fingertip Haptic Sensing of the Visual Environment. *IEEE Journal of Translational Engineering in Health and Medicine* 2: 9 pages. https://doi.org/10.1109/JTEHM.2014.2309343

[80] De-Shuang Huang, Wei Jia, and David Zhang. 2008. Palmprint verification based on principal lines. *Pattern Recognition* 41, 4: 1316–1328. https://doi.org/10.1016/j.patcog.2007.08.016

[81] Andreas Hub, Joachim Diepstraten, and Thomas Ertl. 2003. Design and development of an indoor navigation and object identification system for the blind. *SIGACCESS Access. Comput.*, 77–78: 147–152. https://doi.org/10.1145/1029014.1028657

[82] Alex D Hwang and Eli Peli. 2014. An augmented-reality edge enhancement application for Google Glass. *Optometry and vision science : official publication of the American Academy of Optometry* 91, 8: 1021–30. https://doi.org/10.1097/OPX.0000000000000326

[83] Shahram Izadi, Andrew Davison, Andrew Fitzgibbon, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Dustin Freeman. 2011. KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, 559. https://doi.org/10.1145/2047196.2047270

[84] L Jagannathan and CV Jawahar. 2005. Perspective Correction Methods for

Camera Based Document Analysis. *Proc. First Int. Workshop on Camera-based Document Analysis and Recognition*: 148–154. Retrieved June 21, 2014 from http://imlab.jp/cbdar2005/proceedings/papers/P6.pdf

[85] A.K. Jain and Jianjiang Feng. 2009. Latent Palmprint Matching. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 6: 1032–1047. https://doi.org/10.1109/TPAMI.2008.242

[86] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P Bigham. 2011. Supporting blind photography. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '11), 203–210. https://doi.org/10.1145/2049536.2049573

[87] Kenneth Johnson. 2002. Neural basis of haptic perception. In *Stevens' Handbook of Experimental Psychology: Volume 1: Sensation and Perception* (3rd Editio), Hal Pashler and Steven Yantis (eds.). Wiley Online Library, 537–580.

[88] K. Kanagalakshmi and E. Chandra. 2011. Performance evaluation of filters in noise removal of fingerprint image. In *2011 3rd International Conf. on Electronics Computer Technology*, 117–121. Retrieved March 28, 2016 from http://ieeexplore.ieee.org.proxy-um.researchport.umd.edu/articleDetails.jsp?arnumber=5941572

[89] Shaun K. Kane, Jeffrey P. Bigham, and Jacob O. Wobbrock. 2008. Slide Rule: Making Mobile Touch Screens Accessible to Blind People Using Multi-touch Interaction Techniques. In *Proc. of the ACM SIGACCESS Conference on Computers and Accessibility*, 73–80.

[90] Shaun K Kane, Jeffrey P Bigham, and Jacob O Wobbrock. 2008. Slide rule: making mobile touch screens accessible to blind people using multi-touch interaction techniques. *Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility*: 73–80. Retrieved November 10, 2010 from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Slide+Rule:+Making+Mobile+Touch+Screens+Accessible+to+Blind+People+Using+Multi-Touch+Interaction+Techniques#0

[91] Shaun K Kane, Brian Frey, and Jacob O Wobbrock. 2013. Access Lens: a gesture-based screen reader for real-world documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 347–350. https://doi.org/10.1145/2470654.2470704

[92] Shaun K Kane, Chandrika Jayant, Jacob O Wobbrock, and Richard E Ladner. 2009. Freedom to Roam: A Study of Mobile Device Adoption and Accessibility for People with Visual and Motor Disabilities. In *Proc. ASSETS*

2009, 115. https://doi.org/10.1145/1639642.1639663

[93] Vivek Kanhangad, Ajay Kumar, and David Zhang. 2011. A Unified Framework for Contactless Hand Verification. *IEEE Transactions on Information Forensics and Security* 6, 3: 1014–1027. https://doi.org/10.1109/TIFS.2011.2121062

[94] R Keefer, Yan Liu, and N Bourbakis. 2013. The development and evaluation of an eyes-free interaction model for mobile reading devices. *Human-Machine Systems, IEEE Transactions on* 43, 1: 76–91. https://doi.org/10.1109/TSMCA.2012.2210413

[95] Deborah Kendrick. 2005. From Optacon to Oblivion: the telesensory story. *American Foundation for the Blind AccessWorld Magazine 6*. Retrieved from http://www.afb.org/afbpress/pub.asp?DocID=aw060403

[96] Vinitha Khambadkar and Eelke Folmer. 2013. GIST: A Gestural Interface for Remote Nonvisual Spatial Perception. In *Proceedings of UIST '13* (UIST '13), 301–310. https://doi.org/10.1145/2501988.2502047

[97] Wolf Kienzle and Ken Hinckley. 2014. LightRing. In *Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14*, 157–160. https://doi.org/10.1145/2642918.2647376

[98] knfb Reading Technology Inc. kReader Mobile. Retrieved from http://www.knfbreader.com/products-kreader-mobile.php

[99] knfb Reading Technology Inc. knfb Reader Classic. Retrieved from http://www.knfbreader.com/products-classic.php

[100] Adams Kong, David Zhang, and Mohamed Kamel. 2009. A survey of palmprint recognition. *Pattern Recognition* 42, 7: 1408–1418. https://doi.org/10.1016/j.patcog.2009.01.018

[101] Ig Mo Koo, Kwangmok Jung, Ja Choon Koo, Jae-do Nam, and Young Kwan Lee. 2008. Development of Soft-Actuator-Based Wearable Tactile Display. *IEEE Transactions on Robotics* 24, 3: 549–558. https://doi.org/10.1109/TRO.2008.921561

[102] K M Kramer, D S Hedin, and D J Rolkosky. 2010. Smartphone based face recognition tool for the blind. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 4538–4541. https://doi.org/10.1109/IEMBS.2010.5626038

[103] Sreekar Krishna, Dirk Colbry, John Black, Vineeth Balasubramanian, and Sethuraman Panchanathan. 2008. A systematic requirements analysis and

development of an assistive device to enhance the social interaction of people who are blind or visually impaired. In *Impaired," Workshop on Computer Vision Applications for the Visually Impaired (CVAVI 08), European Conference on Computer Vision ECCV 2008*.

[104] Sreekar Krishna, Greg Little, John Black, and Sethuraman Panchanathan. 2005. A wearable face recognition system for individuals with visual impairments. In *Proc. ASSETS 2005*, 106. https://doi.org/10.1145/1090785.1090806

[105] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 1097–1105. Retrieved June 11, 2017 from http://dl.acm.org/citation.cfm?id=2999257

[106] Moussadek Laadjel, Ahmed Bouridane, Fatih Kurugollu, Omar Nibouche, and WeiQi Yan. 2010. Partial Palmprint Matching Using Invariant Local Minutiae Descriptors. In *Transactions on Data Hiding and Multimedia Security*. 1–17. https://doi.org/10.1007/978-3-642-14298-7_1

[107] Steven Landau and Lesley Wells. 2003. Merging Tactile Sensory Input and Audio Data by Means of The Talking Tactile Tablet. *Proc. Eurographics'03* 2, 60: 414–418.

[108] Maaike Langelaan, Michiel R. de Boer, Ruth M. A. van Nispen, Bill Wouters, Annette C. Moll, and Ger H. M. B. van Rens. 2007. Impact of Visual Impairment on Quality of Life: A Comparison With Quality of Life in the General Population and With Other Chronic Conditions. *Ophthalmic Epidemiology* 14, 3: 119–126. https://doi.org/10.1080/09286580601139212

[109] Gierad Laput, Robert Xiao, Xiang "Anthony" Chen, Scott E. Hudson, and Chris Harrison. 2014. Skin buttons. In *Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14*, 389–394. https://doi.org/10.1145/2642918.2647356

[110] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16*, 321–333. https://doi.org/10.1145/2984511.2984582

[111] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What frustrates screen reader users on the web: a study of 100 blind users. *International Journal of Human-Computer Interaction* 22, 3: 247–269. https://doi.org/10.1080/10447310709336964

[112] Susan J Lederman and Roberta L Klatzky. 1987. Hand movements: A window into haptic object recognition. *Cognitive Psychology* 19, 3: 342–368.

[113] Barbara Leporini, Maria Claudia Buzzi, and Marina Buzzi. 2012. Interacting with mobile devices via VoiceOver: usability and accessibility issues. In *Proceedings of the 24th Australian Computer-Human Interaction Conference on - OzCHI '12*, 339–348. https://doi.org/10.1145/2414536.2414591

[114] Betty Ann Levy. 2001. Text processing: Memory representations mediate fluent reading. *Perspectives on human memory and cognitive aging: Essays in honour of Fergus Craik*: 83–98.

[115] John P. Lewis. 1995. Fast Template Matching. *Vision Interface* 95, 120123: 15–19.

[116] Wei Li, David Zhang, Lei Zhang, Guangming Lu, and Jingqi Yan. 2011. 3-D Palmprint Recognition With Joint Line and Orientation Features. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 2: 274–279. https://doi.org/10.1109/TSMCC.2010.2055849

[117] Rong-Hao Liang, Shu-Yang Lin, Chao-Huai Su, Kai-Yin Cheng, Bing-Yu Chen, and De-Nian Yang. 2011. SonarWatch: Appropriating the Forearm as a Slider Bar. In *SIGGRAPH Asia 2011 Emerging Technologies on - SA '11*, 1–1. https://doi.org/10.1145/2073370.2073374

[118] Soo-Chul Lim, Jungsoon Shin, Seung-Chan Kim, and Joonah Park. 2015. Expansion of Smartwatch Touch Interface from Touchscreen to Around Device Interface Using Infrared Line Image Sensors. *Sensors* 15, 7: 16642–16653. https://doi.org/10.3390/s150716642

[119] Shu-Yang Lin, Chao-Huai Su, Kai-Yin Cheng, Rong-Hao Liang, Tzu-Hao Kuo, and Bing-Yu Chen. 2011. Pub - Point Upon Body: Exploring Eyes-free Interaction and Methods on an Arm. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (UIST '11), 481–488. https://doi.org/10.1145/2047196.2047259

[120] Xu Liu. 2008. A Camera Phone Based Currency Reader for the Visually Impaired. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility* (Assets '08), 305–306. https://doi.org/10.1145/1414471.1414551

[121] Jack M. Loomis, Reginald G. Golledge, and Roberta L. Klatzky. 1998. Navigation System for the Blind: Auditory Display Modes and Guidance. *Presence: Teleoperators and Virtual Environments* 7, 2: 193–203. https://doi.org/10.1162/105474698565677

[122] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2: 91–110.

[123] J. Lucke. 2012. Autonomous cleaning of corrupted scanned documents — A generative modeling approach. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3338–3345. https://doi.org/10.1109/CVPR.2012.6248072

[124] S. O. H. Madgwick, A. J. L. Harrison, and R. Vaidyanathan. 2011. Estimation of IMU and MARG orientation using a gradient descent algorithm. In *2011 IEEE International Conference on Rehabilitation Robotics*, 1–7. https://doi.org/10.1109/ICORR.2011.5975346

[125] Roberto Manduchi. 2012. Mobile Vision as Assistive Technology for the Blind: An Experimental Study. In *Computers Helping People with Special Needs SE - 2*, Klaus Miesenberger, Arthur Karshmer, Petr Penaz and Wolfgang Zagler (eds.). Springer Berlin Heidelberg, 9–16. https://doi.org/10.1007/978-3-642-31534-3_2

[126] Roberto Manduchi and James Coughlan. 2012. (Computer) Vision without Sight. *Communications of the ACM* 55, 1: 96–104. Retrieved November 22, 2013 from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3398697&tool=pmcentrez&rendertype=abstract

[127] Roberto Manduchi and James M Coughlan. 2014. The last meter: blind visual guidance to a target. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*, To appear.

[128] Steve Mann, Jason Huang, Ryan Janzen, Raymond Lo, Valmiki Rampersad, Alexander Chen, and Taqveer Doha. 2011. Blind navigation with a wearable range camera and vibrotactile helmet. In *Proceedings of the 19th ACM international conference on Multimedia - MM '11*, 1325. Retrieved August 24, 2015 from http://dl.acm.org/citation.cfm?id=2072298.2072005

[129] Heinrich Braun Martin Riedmiller. 1992. RPROP - A Fast Adaptive Learning Algorithm. In *Proc. of ISCIS VII*.

[130] R W Massof and D L Rickman. 1992. Obstacles encountered in the development of the low vision enhancement system. *Optometry and vision science : official publication of the American Academy of Optometry* 69, 1: 32–41. Retrieved March 26, 2018 from http://www.ncbi.nlm.nih.gov/pubmed/1371334

[131] Denys J. C. Matthies, Simon T. Perrault, Bodo Urban, and Shengdong Zhao.

2015. Botential: Localizing On-Body Gestures by Measuring Electrical Signatures on the Human Skin. In *Proc. MobileHCI 2015*, 207–216. https://doi.org/10.1145/2785830.2785859

[132] Stefano Mattoccia and Paolo Macrı'. 2015. 3D Glasses as Mobility Aid for Visually Impaired People. . Springer, Cham, 539–554. https://doi.org/10.1007/978-3-319-16199-0_38

[133] W.W. Mayol-Cuevas, B.J. Tordoff, and D.W. Murray. 2009. On the Choice and Placement of Wearable Vision Sensors. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 39, 2: 414–425. https://doi.org/10.1109/TSMCA.2008.2010848

[134] David McGookin, Stephen Brewster, and WeiWei Jiang. 2008. Investigating touchscreen accessibility for people with visual impairments. In *Proceedings of the Nordic Conference on Human-Computer Interaction (NordiCHI'08)*, 298–307. Retrieved September 18, 2013 from http://dl.acm.org/citation.cfm?id=1463193

[135] Alexander J. Medeiros, Lee Stearns, Leah Findlater, Chuan Chen, and Jon E. Froehlich. 2017. Recognizing Clothing Colors and Visual Textures Using a Finger-Mounted Camera: An Initial Investigation. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '17*, 393–394. https://doi.org/10.1145/3132525.3134805

[136] Abdallah Meraoumia, Salim Chitroub, and Ahmed Bouridane. 2011. Fusion of Finger-Knuckle-Print and Palmprint for an Efficient Multi-Biometric System of Person Recognition. In *2011 IEEE International Conference on Communications (ICC)*, 1–5. https://doi.org/10.1109/icc.2011.5962661

[137] David Merrill and Pattie Maes. 2007. Augmenting Looking, Pointing and Reaching Gestures to Enhance the Searching and Browsing of Physical Objects. In *Pervasive Computing*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–18. https://doi.org/10.1007/978-3-540-72037-9_1

[138] Susanna Millar. 2003. *Reading by touch*. Routledge. Retrieved November 7, 2014 from http://books.google.com/books?hl=en&lr=&id=maCJAgAAQBAJ&pgis=1

[139] Pranav Mistry and Pattie Maes. 2009. SixthSense: A wearable gestural interface. In *Proceedings of ACM SIGGRAPH Asia*, Article No. 11.

[140] Elad Moisseiev, Mark J. Mannis, Liu CJ, Berger S, Renieri G, Taylor J, Moshtael H, Crossland MD, Virgili G, and Peterson RC. 2016. Evaluation of a Portable Artificial Vision Device Among Patients With Low Vision. *JAMA*

Ophthalmology 134, 7: 748. https://doi.org/10.1001/jamaophthalmol.2016.1000

[141] Aythami Morales, Miguel A. Ferrer, and Ajay Kumar. 2010. Improved palmprint authentication using contactless imaging. In *IEEE Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, 1–6. https://doi.org/10.1109/BTAS.2010.5634472

[142] S Mori, C Y Suen, and K Yamamoto. 1992. Historical review of OCR research and development. *Proceedings of the IEEE* 80, 7: 1029–1058. https://doi.org/10.1109/5.156468

[143] Dan Morris, T Scott Saponas, and Desney Tan. 2010. Emerging input technologies for always-available mobile interaction. *Foundations and Trends in Human--Computer Interaction* 4, 4: 245–316.

[144] Randall Munroe. 2010. Color Survey Results. Retrieved from https://blog.xkcd.com/2010/05/03/color-survey-results/

[145] Suranga Nanayakkara, Roy Shilkrot, and Pattie Maes. 2012. EyeRing: A Finger-worn Assistant. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, 1961–1966. https://doi.org/10.1145/2212776.2223736

[146] Suranga Nanayakkara, Roy Shilkrot, Kian Peen Yeo, and Pattie Maes. 2013. EyeRing: A Finger-worn Input Device for Seamless Interactions with Our Surroundings. In *Proc. AH 2013*, 13–20. https://doi.org/10.1145/2459236.2459240

[147] National Center for Health Statistics. 2012. *National Health Interview Survey 2012 Data Release*. Retrieved from http://www.cdc.gov/nchs/nhis/nhis_2012_data_release.htm

[148] National Federation of the Blind Jernigan Institute. 2009. The braille literacy crisis in america: facing the truth, reversing the trend, empowering the blind. Retrieved from https://nfb.org/images/nfb/publications/bm/bm09/bm0905/bm090504.htm

[149] J Farley Norman and Ashley N Bartholomew. 2011. Blindness enhances tactile acuity and haptic 3-D shape discrimination. *Attention, Perception, & Psychophysics* 73, 7: 2323–2331. https://doi.org/10.3758/s13414-011-0160-4

[150] Masa Ogata and Michita Imai. 2015. SkinWatch: Skin Gesture Interaction for Smart Watch. In *Proceedings of the 6th Augmented Human International Conference on - AH '15*, 21–24. https://doi.org/10.1145/2735711.2735830

[151] Masa Ogata, Yuta Sugiura, Yasutoshi Makino, Masahiko Inami, and Michita

Imai. 2013. SenSkin: Adapting Skin as a Soft Interface. In *Proceedings of the 26th annual ACM symposium on User interface software and technology - UIST '13*, 539–544. https://doi.org/10.1145/2501988.2502039

[152] Masa Ogata, Yuta Sugiura, Yasutoshi Makino, Masahiko Inami, and Michita Imai. 2014. Augmenting a Wearable Display with Skin Surface as an Expanded Input Area. In *Design, User Experience, and Usability. User Experience Design for Diverse Interaction Platforms and Environments*, Aaron Marcus (ed.). Springer International Publishing, Cham, 606–614. https://doi.org/10.1007/978-3-319-07626-3_57

[153] Uran Oh and Leah Findlater. 2014. Design of and subjective response to on-body input for people with visual impairments. In *Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '14)*, 8 pages.

[154] Uran Oh and Leah Findlater. 2015. A Performance Comparison of On-Hand versus On-Phone Non-Visual Input by Blind and Sighted Users. *ACM Transactions on Accessible Computing (TACCESS)* 7, 4: 14.

[155] Uran Oh, Shaun K Kane, and Leah Findlater. 2013. Follow That Sound: Using Sonification and Corrective Verbal Feedback to Teach Touchscreen Gestures. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '13), 13:1--13:8. https://doi.org/10.1145/2513383.2513455

[156] Uran Oh, Lee Stearns, Alisha Pradhan, Jon E. Froehlich, and Leah Findlater. 2017. Investigating Microinteractions for People with Visual Impairments and the Potential Role of On-Body Interaction. In *ASSETS 2017*, TO APPEAR.

[157] T. Ojala, M. Pietikainen, and T. Maenpaa. 2002. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24, 7: 971–987. https://doi.org/10.1109/TPAMI.2002.1017623

[158] Goh Kah Ong Michael, Tee Connie, and Andrew Beng Jin Teoh. 2008. Touch-less palm print biometrics: Novel design and implementation. *Image and Vision Computing* 26, 12: 1551–1560. https://doi.org/10.1016/j.imavis.2008.06.010

[159] OrCam Technologies Ltd. OrCam - See for Yourself. Retrieved from http://www.orcam.com/

[160] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10: 1345–1359.

https://doi.org/10.1109/TKDE.2009.191

[161] Peter Parente and Gary Bishop. 2003. BATS: The Blind Audio Tactile Mapping System. *Proceedings of ACM South Eastern Conference*. Retrieved August 13, 2015 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.6189

[162] Anil Singh Parihar, Amoiy Kumar, Om Prakash Verma, Ankita Gupta, Prerana Mukherjee, and Deepika Vatsa. 2013. Point based features for contact-less palmprint images. In *2013 IEEE International Conference on Technologies for Homeland Security (HST)*, 165–170. https://doi.org/10.1109/THS.2013.6698994

[163] Marcin Pazio, Maciej Niedzwiecki, Ryszard Kowalik, and Jacek Lebiedz. 2007. Text detection system for the blind. In *15th European Signal Processing Conference (EUSIPCO 2007)*, 272–276. Retrieved from http://www.eurasip.org/Proceedings/Eusipco/Eusipco2007/Papers/a2l-f02.pdf

[164] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. 2010. Improving the Fisher Kernel for Large-Scale Image Classification. . Springer, Berlin, Heidelberg, 143–156. https://doi.org/10.1007/978-3-642-15561-1_11

[165] John C. Platt. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers* 10, 3: 61–74.

[166] Benjamin Poppinga, Charlotte Magnusson, Martin Pielot, and Kirsten Rassmus-Gröhn. 2011. TouchOver map: audio-tactile exploration of interactive maps. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI '11*, 545. Retrieved August 13, 2015 from http://dl.acm.org.proxy-um.researchport.umd.edu/citation.cfm?id=2037373.2037458

[167] The Princeton Review. 2000. *ITBS preparation and practice workbook, Glencoe language arts, grade 8*. McGraw-Hill/Glencoe, New York, NY.

[168] Halley Profita, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun K. Kane. 2016. The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use. In *Proc. CHI 2016*, 4884–4895. https://doi.org/10.1145/2858036.2858130

[169] Halley P. Profita, James Clawson, Scott Gilliland, Clint Zeagler, Thad Starner, Jim Budd, and Ellen Yi-Luen Do. 2013. Don't mind me touching my wrist. In *Proceedings of the 17th annual international symposium on International symposium on wearable computers - ISWC '13*, 89. Retrieved February 27,

2015 from http://dl.acm.org/citation.cfm?id=2493988.2494331

[170] Shrinivas Pundlik, Huaqi Yi, Rui Liu, Eli Peli, and Gang Luo. 2016. Magnifying Smartphone Screen using Google Glass for Low-Vision Users. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*: 1–1. https://doi.org/10.1109/TNSRE.2016.2546062

[171] L. Ran, S. Helal, and S.E. Moore. 2004. Drishti: an integrated indoor/outdoor blind navigation system and service. In *Second IEEE Annual Conference on Pervasive Computing and Communications, 2004. Proceedings of the*, 23–30. https://doi.org/10.1109/PERCOM.2004.1276842

[172] Shanaka Ransiri and Suranga Nanayakkara. 2013. SmartFinger: An Augmented Finger As a Seamless "Channel" Between Digital and Physical Objects. In *Proceedings of the 4th Augmented Human International Conference* (AH '13), 5–8. https://doi.org/10.1145/2459236.2459238

[173] J. Rekimoto. 2001. GestureWrist and GesturePad: unobtrusive wearable interaction devices. In *Proceedings Fifth International Symposium on Wearable Computers*, 21–27. Retrieved January 20, 2014 from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=962092

[174] Jun Rekimoto. 2002. SmartSkin: an infrastructure for freehand manipulation on interactive surfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02*, 113. https://doi.org/10.1145/503376.503397

[175] Julie Rico and Stephen Brewster. 2009. Gestures all around us: user differences in social acceptability perceptions of gesture based interfaces. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '09*, 64. https://doi.org/10.1145/1613858.1613936

[176] Julie Rico and Stephen Brewster. 2010. Usable gestures for mobile interfaces. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, 887. https://doi.org/10.1145/1753326.1753458

[177] Mikko J. Rissanen, Samantha Vu, Owen Noel Newton Fernando, Natalie Pang, and Schubert Foo. 2013. Subtle, Natural and Socially Acceptable Interaction Techniques for Ringterfaces — Finger-Ring Shaped User Interfaces. In *Distributed, Ambient, and Pervasive Interactions SE - 6*, Norbert Streitz and Constantine Stephanidis (eds.). Springer Berlin Heidelberg, 52–61. https://doi.org/10.1007/978-3-642-39351-8_6

[178] Alberto Rodríguez, J. Javier Yebes, Pablo Alcantarilla, Luis Bergasa, Javier

Almazán, and Andrés Cela. 2012. Assisting the Visually Impaired: Obstacle Detection and Warning System by Acoustic Feedback. *Sensors* 12, 12: 17476–17496. https://doi.org/10.3390/s121217476

[179] E. Rosten and T. Drummond. 2005. Fusing points and lines for high performance tracking. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 1508–1515 Vol. 2. https://doi.org/10.1109/ICCV.2005.104

[180] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3: 211–252. https://doi.org/10.1007/s11263-015-0816-y

[181] R. Safabakhsh and S. Khadivi. 2000. Document skew detection using minimum-area bounding rectangle. In *Proceedings International Conference on Information Technology: Coding and Computing (Cat. No.PR00540)*, 253–258. https://doi.org/10.1109/ITCC.2000.844226

[182] T. Scott Saponas. 2010. Supporting everyday activities through always-available mobile computing. University of Washington. Retrieved from http://research.microsoft.com/en-us/um/people/ssaponas/publications/Saponas-Dissertation.pdf

[183] T Scott Saponas, Desney S Tan, Dan Morris, Ravin Balakrishnan, Jim Turner, and James A Landay. 2009. Enabling Always-available Input with Muscle-computer Interfaces. In *Proceedings of the 22Nd Annual ACM Symposium on User Interface Software and Technology* (UIST '09), 167–176. https://doi.org/10.1145/1622176.1622208

[184] Munehiko Sato, Ivan Poupyrev, and Chris Harrison. 2012. Touché: enhancing touch interaction on humans, screens, liquids, and everyday objects. *Proceedings of CHI '12*, c: 483–492. https://doi.org/10.1145/2207676.2207743

[185] Douglas. Schuler and Aki. Namioka. 1993. *Participatory design : principles and practices*. L. Erlbaum Associates. Retrieved April 16, 2018 from https://dl.acm.org/citation.cfm?id=563076

[186] Lionel Prevost Shehzad Muhammad Hanif. 2007. Texture based Text Detection in Natural Scene Images-A Help to Blind and Visually Impaired Persons. In *Proceedings of CVHI 2007*. Retrieved June 2, 2016 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.142.8931

[187] Huiying Shen and James M. Coughlan. 2012. Towards a real-time system for

finding and reading signs for visually impaired users. In *Computers Helping People with Special Needs SE - 7*, Klaus Miesenberger, Arthur Karshmer, Petr Penaz and Wolfgang Zagler (eds.). Springer Berlin Heidelberg, 41–47. https://doi.org/10.1007/978-3-642-31534-3_7

[188] Roy Shilkrot, Jochen Huber, Roger Boldu, Pattie Maes, and Suranga Nanayakkara. 2018. FingerReader: A Finger-Worn Assistive Augmentation. . Springer, Singapore, 151–175. https://doi.org/10.1007/978-981-10-6404-3_9

[189] Roy Shilkrot, Jochen Huber, Connie Liu, Pattie Maes, and Nanayakkara Suranga Chandima. 2014. FingerReader: a wearable device to support text reading on the go. *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, Vi: 2359–2364. https://doi.org/10.1145/2559206.2581220

[190] Roy Shilkrot, Jochen Huber, Wong Meng Ee, Pattie Maes, and Suranga Chandima Nanayakkara. 2015. FingerReader: A Wearable Device to Explore Printed Text on the Go. In *Proc. CHI 2015*, 2363–2372. https://doi.org/10.1145/2702123.2702421

[191] Roy Shilkrot, Jochen Huber, Jürgen Steimle, Suranga Nanayakkara, and Pattie Maes. 2015. Digital Digits: A Comprehensive Survey of Finger Augmentation Devices. *ACM Computing Surveys* 48, 2: 1–29. https://doi.org/10.1145/2828993

[192] Kristen Shinohara and Jacob O Wobbrock. 2011. In the Shadow of Misperception: Assistive Technology Use and Social Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 705–714. https://doi.org/10.1145/1978942.1979044

[193] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. Retrieved June 12, 2018 from http://arxiv.org/abs/1409.1556

[194] Laurindo de Sousa Britto Neto, Vanessa Regina Margareth Lima Maike, Fernando Luiz Koch, Maria Cecília Calani Baranauskas, Anderson de Rezende Rocha, and Siome Klein Goldenstein. 2015. A Wearable Face Recognition System Built into a Smartwatch and the Blind and Low Vision Users. In *Proc. ASSETS 2015*, 515–528. https://doi.org/10.1007/978-3-319-29133-8_25

[195] Srinath Sridhar, Anders Markussen, Antti Oulasvirta, Christian Theobalt, and Sebastian Boring. 2017. WatchSense: On- and Above-Skin Input Sensing through a Wearable Depth Sensor. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 3891–3902. https://doi.org/10.1145/3025453.3026005

[196] T. Starner, J. Auxier, D. Ashbrook, and M. Gandy. The gesture pendant: a self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. In *Digest of Papers. Fourth International Symposium on Wearable Computers*, 87–94. https://doi.org/10.1109/ISWC.2000.888469

[197] Lee Stearns, Victor DeSouza, Jessica Yin, Leah Findlater, and Jon E. Froehlich. 2017. Augmented Reality Magnification for Low Vision Users with the Microsoft Hololens and a Finger-Worn Camera. In *Proc. ASSETS 2017*, 361–362. https://doi.org/10.1145/3132525.3134812

[198] Lee Stearns, Ruofei Du, Uran Oh, Catherine Jou, Leah Findlater, David A. Ross, and Jon E. Froehlich. 2016. Evaluating Haptic and Auditory Directional Guidance to Assist Blind People in Reading Printed Text Using Finger-Mounted Cameras. *ACM Transactions on Accessible Computing* 9, 1: 1–38. https://doi.org/10.1145/2914793

[199] Lee Stearns, Ruofei Du, Uran Oh, Yumeng Wang, Rama Chellappa, Leah Findlater, and Jon E. Froehlich. 2014. The Design and Preliminary Evaluation of a Finger-Mounted Camera and Feedback System to Enable Reading of Printed Text for the Blind. *Workshop on Assistive Computer Vision and Robotics (ACVR'14) in Conjunction with the European Conference on Computer Vision (ECCV'14)*: 615--631. https://doi.org/10.1007/978-3-319-16199-0_43

[200] Lee Stearns, Leah Findlater, and Jon E. Froehlich. 2018. Design of an Augmented Reality Magnification Aid for Low Vision Users. In *Proc. ASSETS 2018 (To Appear)*.

[201] Lee Stearns, Leah Findlater, and Jon E. Froehlich. 2018. Applying Transfer Learning to Recognize Clothing Patterns Using a Finger-Mounted Camera. In *Proc. ASSETS 2018 (To Appear)*.

[202] Lee Stearns, Uran Oh, Bridget J. Cheng, Leah Findlater, David Ross, Rama Chellappa, and Jon E. Froehlich. 2016. Localization of skin features on the hand and wrist from small image patches. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 1003–1010. https://doi.org/10.1109/ICPR.2016.7899767

[203] Lee Stearns, Uran Oh, Leah Findlater, and Jon E. Froehlich. 2018. TouchCam: Realtime Recognition of Location-Specific On-Body Gestures to Support Users with Visual Impairments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4: 1–23. https://doi.org/10.1145/3161416

[204] Jeremi Sudol, Orang Dialameh, Chuck Blanchard, and Tim Dorcey. 2010. Looktel—A comprehensive platform for computer-aided visual assistance. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 73–80. https://doi.org/10.1109/CVPRW.2010.5543725

[205] Catherine A Sugar and Gareth M James. 2003. Finding the Number of Clusters in a Dataset. *Journal of the American Statistical Association* 98, 463: 750–763. https://doi.org/10.1198/016214503000000666

[206] Emi Tamaki, Takashi Miyaki, and Jun Rekimoto. 2009. Brainy Hand: an earworn hand gesture interaction device. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems - CHI EA '09*, 4255. https://doi.org/10.1145/1520340.1520649

[207] Makoto Tanaka and Hideaki Goto. 2008. Text-tracking wearable camera system for visually-impaired people. In *2008 19th International Conference on Pattern Recognition*, 1–4. Retrieved July 6, 2015 from http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4761548

[208] Brandon Taylor, Anind Dey, Dan Siewiorek, and Asim Smailagic. 2016. Customizable 3D Printed Tactile Maps as Interactive Overlays. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '16*, 71–79. https://doi.org/10.1145/2982142.2982167

[209] Yingli Tian and Shuai Yuan. 2010. Clothes Matching for Blind and Color Blind People. In *Computers Helping People with Special Needs SE - 48*, Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler and Arthur Karshmer (eds.). Springer Berlin Heidelberg, 324–331. https://doi.org/10.1007/978-3-642-14100-3_48

[210] J. L. Toennies, J. Burgner, T. J. Withrow, and R. J. Webster. 2011. Toward haptic/aural touchscreen display of graphical mathematics for the education of blind students. In *2011 IEEE World Haptics Conference*, 373–378. Retrieved August 13, 2015 from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5945515

[211] Carlo Tomasi and Takeo Kanade. 1991. *Detection and Tracking of Point Features*. Carnegie Mellon University Technical Report CMU-CS-91-132.

[212] Sylvie Treuillet, Eric Royer, Thierry Chateau, Michel Dhome, and Jean-Marc Lavest. 2006. Body Mounted Vision System For Visually Impaired Outdoor And Indoor Wayfindind Assistance. none. Retrieved July 12, 2018 from https://hal.archives-ouvertes.fr/hal-00167383/

[213] Marynel Vázquez and Aaron Steinfeld. 2012. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '12*, 95. https://doi.org/10.1145/2384916.2384934

[214] R. Velazquez, E.E. Pissaloux, M. Hafez, and J. Szewczyk. 2008. Tactile Rendering With Shape-Memory-Alloy Pin-Matrix. *IEEE Transactions on Instrumentation and Measurement* 57, 5: 1051–1057. https://doi.org/10.1109/TIM.2007.913768

[215] Ramiro Velázquez. 2010. Wearable Assistive Devices for the Blind. In *Wearable and Autonomous Systems*, 331–349.

[216] Wai Kin Kong and D. Zhang. 2002. Palmprint texture analysis based on low-resolution images for personal authentication. In *Proc. Pattern Recognition '02*, 807–810. https://doi.org/10.1109/ICPR.2002.1048142

[217] Michael Waisbourd, Osama Ahmed, Linda Siam, Marlene R Moster, Lisa A Hark, and L Jay Katz. 2015. The Impact of a Novel Artificial Vision Device (OrCam) on the Quality of Life of Patients with End-Stage Glaucoma. *Investigative Ophthalmology & Visual Science* 56, 7: 519–519.

[218] Cheng-Yao Wang, Min-Chieh Hsiu, Po-Tsung Chiu, Chiao-Hui Chang, Liwei Chan, Bing-Yu Chen, and Mike Y. Chen. 2015. PalmGesture: Using Palms as Gesture Interfaces for Eyes-free Input. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '15*, 217–226. https://doi.org/10.1145/2785830.2785885

[219] Kai Wang, B Babenko, and S Belongie. 2011. End-to-end scene text recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 1457–1464. https://doi.org/10.1109/ICCV.2011.6126402

[220] Martin Weigel, Tong Lu, Gilles Bailly, Antti Oulasvirta, Carmel Majidi, and Jürgen Steimle. 2015. iSkin: Flexible, Stretchable and Visually Customizable On-Body Touch Sensors for Mobile Computing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 2991–3000. https://doi.org/10.1145/2702123.2702391

[221] Michele A Williams, Callie Neylan, and Amy Hurst. 2013. Preliminary Investigation of the Limitations Fashion Presents to Those with Vision Impairments. *Fashion Practice: The Journal of Design, Creative Process & the Fashion* 5, 1: 81–106. https://doi.org/doi:10.2752/175693813X13559997788808

[222] Andrew D. Wilson and Hrvoje Benko. 2010. Combining multiple depth

cameras and projectors for interactions on, above and between surfaces. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10*, 273. https://doi.org/10.1145/1866029.1866073

[223] Andrew D. Wilson and Andrew D. 2010. Using a depth camera as a touch sensor. In *ACM International Conference on Interactive Tabletops and Surfaces - ITS '10*, 69. https://doi.org/10.1145/1936652.1936665

[224] Jiahui Wu, Gang Pan, Daqing Zhang, Guande Qi, and Shijian Li. 2009. Gesture Recognition with a 3-D Accelerometer. In *Ubiquitous intelligence and computing (UIC 2009)*, 25–38. https://doi.org/10.1007/978-3-642-02830-4_4

[225] Xiangqian Wu, Qiushi Zhao, and Wei Bu. 2014. A SIFT-based contactless palmprint verification approach using iterative RANSAC and local palmprint descriptors. *Pattern Recognition* 47, 10: 3314–3326. https://doi.org/10.1016/j.patcog.2014.04.008

[226] Robert Xiao, Scott Hudson, and Chris Harrison. 2016. DIRECT: Making Touch Tracking on Ordinary Surfaces Practical with Hybrid Depth-Infrared Sensing. In *Proceedings of the 2016 ACM on Interactive Surfaces and Spaces - ISS '16*, 85–94. https://doi.org/10.1145/2992154.2992173

[227] Xiaodong Yang, Shuai Yuan, and YingLi Tian. 2014. Assistive Clothing Pattern Recognition for Visually Impaired People. *IEEE Transactions on Human-Machine Systems* 44, 2: 234–243. https://doi.org/10.1109/THMS.2014.2302814

[228] Xing-Dong Yang, Tovi Grossman, Daniel Wigdor, and George Fitzmaurice. 2012. Magic finger: always-available input through finger instrumentation. In *Proceedings of UIST '12* (UIST '12), 147–156. https://doi.org/10.1145/2380116.2380137

[229] Hanlu Ye, Meethu Malu, Uran Oh, and Leah Findlater. 2014. Current and future mobile and wearable device use by people with visual impairments. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2014)*, 10 pages.

[230] Chucai Yi, YingLi Tian, and A Arditi. 2014. Portable Camera-Based Assistive Text and Product Label Reading From Hand-Held Objects for Blind Persons. *Mechatronics, IEEE/ASME Transactions on* 19, 3: 808–817. https://doi.org/10.1109/TMECH.2013.2261083

[231] Shuai Yuan, YingLi Tian, and Aries Arditi. 2011. Clothing matching for visually impaired persons. *Technology and Disability* 23, 2: 75–85. https://doi.org/10.3233/TAD-2011-0313

[232] Vadim Zaliva. 2012. Horizontal Perspective Correction in Text Images. Retrieved from http://notbrainsurgery.livejournal.com/40465.html

[233] Ye Ze, Chucai Yi, and YingLi Tian. 2013. Reading labels of cylinder objects for blind persons. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, 1–6. https://doi.org/10.1109/ICME.2013.6607632

[234] Yang Zhang, Junhan Zhou, Gierad Laput, and Chris Harrison. 2016. SkinTrack: Using the Body as an Electrical Waveguide for Continuous Finger Tracking on the Skin. In *Proc. CHI 2016*, 1491–1503. https://doi.org/10.1145/2858036.2858082

[235] Yuhang Zhao, Sarit Szpiro, and Shiri Azenkot. 2015. ForeSee: A Customizable Head-Mounted Vision Enhancement System for People with Low Vision. In *Proc. ASSETS 2015*, 239–249. https://doi.org/10.1145/2700648.2809865

[236] Zhi Li, Guizhong Liu, Yang Yang, and Junyong You. 2012. Scale- and Rotation-Invariant Local Binary Pattern Using Scale-Adaptive Texton and Subuniform-Based Circular Shift. *IEEE Transactions on Image Processing* 21, 4: 2130–2140. https://doi.org/10.1109/TIP.2011.2173697

[237] Annuska Zolyomi, Anushree Shukla, and Jaime Snyder. 2017. Technology-Mediated Sight: A Case Study of Early Adopters of a Low Vision Assistive Technology. *Proc. ASSETS 2017*: 220–229. https://doi.org/10.1145/3132525.3132552

[238] eSight Glasses. Retrieved from https://www.esighteyewear.com/technology

[239] NuEyes Pro. Retrieved from https://nueyes.com/nueyes-pro/

[240] IrisVision. Retrieved from https://irisvision.com/

[241] Microsoft Hololens. Retrieved from https://www.microsoft.com/en-us/hololens

[242] Microsoft Seeing AI. Retrieved from https://www.microsoft.com/en-us/seeing-ai/

[243] Ruby Handheld Video Magnifier. Retrieved from http://www.freedomscientific.com/Products/LowVision/Ruby

[244] Optelec Compact+ HD. Retrieved from https://us.optelec.com/products/compact-hd.html

[245] Oculus Rift. Retrieved from https://www.oculus.com/rift/

[246] Apple iOS ARKit. Retrieved from https://developer.apple.com/arkit/