Modern Sensory Substitution for Vision in Dynamic Environments

by

Bijan Fakhri

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2020 by the
Graduate Supervisory Committee:

Sethuraman Panchanathan, Chair
Troy L. McDaniel
Hemanth Venkateswara
Heni Amor

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

Societal infrastructure is built with vision at the forefront of daily life. For those with severe visual impairments, this creates countless barriers to the participation and enjoyment of life's opportunities. Technological progress has been both a blessing and a curse in this regard. Digital text together with screen readers and refreshable Braille displays have made whole libraries readily accessible and rideshare tech has made independent mobility more attainable. Simultaneously, screen-based interactions and experiences have only grown in pervasiveness and importance, precluding many of those with visual impairments.

Sensory Substituion, the process of substituting an unavailable modality with another one, has shown promise as an alternative to accomodation, but in recent years meaningful strides in Sensory Substitution for vision have declined in frequency. Given recent advances in Computer Vision, this stagnation is especially disconcerting. Designing Sensory Substitution Devices (SSDs) for vision for use in interactive settings that leverage modern Computer Vision techniques presents a variety of challenges including perceptual bandwidth, human-computer-interaction, and person-centered machine learning considerations. To surmount these barriers an approach called Personal Foveated Haptic Gaze (PFHG), is introduced. PFHG consists of two primary components: a human visual system inspired interaction paradigm that is intuitive and flexible enough to generalize to a variety of applications called Foveated Haptic Gaze (FHG), and a person-centered learning component to address the expressivity limitations of most SSDs. This component is called One-Shot Object Detection by Data Augmentation (1SODDA), a one-shot object detection approach that allows a user to specify the objects they are interested in locating visually and with minimal effort realizing an object detection model that does so effectively.

The Personal Foveated Haptic Gaze framework was realized in a virtual and real-world application: playing a 3D, interactive, first person video game (DOOM) and finding user-specified real-world objects. User study results found Foveated Haptic Gaze to be an effective and intuitive interface for interacting with dynamic visual world using solely haptics. Additionally, 1SODDA achieves competitive performance among few-shot object detection methods and high-framerate many-shot object detectors. The combination of which paves the way for modern Sensory Substitution Devices for vision.

DEDICATION

To my mom, dad, and brother, whose compassion, resolve, and affinity for the unknown continue to inspire me. To Michelle, whose enterprising passion emboldens my spirit.

ACKNOWLEDGMENTS

I would firstly like to thank my mentor, advisor and committee chair, Dr. Sethuraman (Panch) Panchanathan. His guidance, encouragement, and confidence in me have continued to remind me of the possibilities that lay ahead. Also, the support of my committee members Dr. Troy McDaniel, Dr. Hemanth Venkateswara, and Dr. Heni Amor was critical to my success as a scholar and professional.

I would also like to thank Abhik Chowdhury and Shashank Sharma whose shared love of hardware and ingenuity made the rapid prototyping of novel interfaces a pleasure.

I would also like to thank the CUbiC family for their comradery and their spirit of intellectual freedom. I also want to thank Jay Klein and the APAcT-IGERT group, whose mission and variety of perspectives challenged me and enriched my academic journey.

Lastly, I would like to thank Zsolt Kira at the Georgia Tech Research Institute, Heni Amor at the Interactive Robotics Lab at ASU, and the Smart Products team at P&G for providing a creative space outside of my home lab.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Worldwide, there are over 200 million people who have some kind of vision impairment, Bourne *et al.* (2017). In the United States alone, there are 7.3 million adults with blindness according to the Cornell University's Employment and Disability Institute, Nations (1990). These individuals face a disproportionate burden in participating in the activities of daily life due to social, societal, and infrastructural barriers. Infrastructural barries are faced early in age, as almost 90% of blind children are not taught Braille and thus do not receive access to fundamental reading and writing education and an estimated 50% of high school students who are blind drop out before graduation, Jernigan Institute (2009). These educational barriers, along with the associated stigma, reduce opportunity early on. Employment is another aspect presenting barriers to people with blindness. The National Federation of the Blind (NFB) estimates that the unemployment rate for people with blindness is greater than 70%, Jernigan Institute (2009). In 2016, the NFB reported that 27.7% of blind individuals in the United States live under the poverty line, NFB (2017), more than double the national poverty rate of 12.7 %, Semega *et al.* (2017). While the Americans with Disabilities Act of 1990 made great strides towards equaility, the data shows that an overwhelming disparity in access to employment still exists. Additionally, the increasing ubiquity of technology has in ways exacerbated the inequity. A survey of teachers of students who are blind found that they spent the majority of their time instructing on how to become a proficient user of assistive technology, Thurlow *et al.* (2001). As computing and technology have become integral to full participation in society, the reliance on screen-based interfaces has increased. Efforts

to guide the development of more inclusive environments, such as Universal Design, Rose *et al.* (2005), show promise but its principles have yet to see widespread adoption in the realm of technology.

Apart from efforts to design more inclusively, technological innovations have made significant improvements in the lives of people with blindness. Most of these advancements can be categorized as methods for delivering visual information to the user using alternate means. The most salient example of this is the Braille system, invented by Louis Braille in 1829, which after standardization brought the written word to many people with blindness. The cane is an even older example of such a tool but used for navigation. Both of these technologies utilize Sensory Substitution as their mode of operation. Sensory Substitution (SS) is the process of delivering a signal from the domain of one sensory modality to an alternative sensory modality. The idea was introduced and popularized by Dr. Paul Bach-y-Rita in the 1960s. While instances of Sensory Substitution predate Dr. Bach-y-Rita's work by centuries, his early work on vision substitution made waves in the field of neuroscience. This work paired a blind individual with a video camera and a large haptic interface on a dental chair which resulted in the blind individual being able to distinguish objects at a distance using the system.

Any device which makes use of Sensory Substitution to function is called a Sensory Substitution Device (SSD). SSDs are not limited to substituting for vision, nor are they limitted to using haptics as a target modality. Researchers are the Eagleman lab, for example, were successful in substituting hearing with haptics, Eagleman (2014) and Novich (2015). Researchers at the Neural Rehabilitation Engineering Laboratory used audition to substitute for vision with their device, the Prosthesis for Substitution of Vision with Audition (PSVA), Capelle *et al.* (1998).

Of all the possible target modalities, haptics has inherent advantages for assistive technology. The skin is the largest organ on the body, giving the organ versatility when it comes to design. The plentiful real-estate of the skin allows haptic SSDs to be designed to impart minimal obstruction to other crucial functions of the senses. Haptic actuators can be placed on surfaces such as the back, upper arms, or waistline, locations that are not often used in day-to-day activity. This is not the case with audition as a modality, as placing anything over the ears obstructs the entirety of hearing. Additionally, modern attention models of the brain partition perceptual bandwidth by modality: the relatively unnused nature of the skin as a sensory organ represents unused perceptual bandwidth via the "Modality Effect". The skin though does exhibit some limitations that have hindered its adoption as the defacto target for general-purpose vision substitution, Spence (2014). This is evidenced by the fact that few haptic SSDs for vision are in use for interactive environments. While screen-readers and refreshable Braille displays have enjoyed growing success, WebAIM (Web Accessibility In Mind) (2015), these modern SSDs are still limited to non-interactive applications.

Most daily activities are interactive in nature. From finding one's misplaced keys to playing a game with friends and family, these activities feature dynamic environments that change with respect to one's actions. These activities require exploration and realtime feedback. Modern haptic SSDs have yet to meet the demands for use in interactive scenarios. In this last decade progress in the field of Computer Vision, especially with the rise of Deep Learning, ellicited hope for a great leap in assistive technology for vision. So far, these solutions have been sparse and of limited use in interactive scenarios, mostly hindered not by the challenges of Computer Vision but by usability limitations. To truly capitalize on the advancements in Computer Vision and develop haptic SSDs that are both general and useful in interactive scenarios, a

more intuitive interface and Computer Vision methods capable of adapting to a user's specific needs are called for. This dissertation introduces Person Centered Foveated Haptic Gaze, an approach addressing this call, paving the way for modern Sensory Substitution for vision.

This dissertation is organized in the following manner. Chapter 2 provides background with regard to the sensory processes of vision and touch, followed by a discussion on the field of Sensory Substitution and Sensory Substitution Devices. Chapter 3 discusses related works with respect to Haptic Sensory Substitution in a variety of application domains. This chapter concludes with the focus of this dissertation: Sensory Substitution for interactive applications in the era of Artificial Intelligence. Chapter 4 explores developments in Deep Reinforcement Learning as applied to visual navigation and their potential for use in assistive technology applications. Chapter 5 introduces a standardized device and protocol called the Low Resolution Haptic Interface (LRHI) for communicating spatial information haptically. This chapter also discusses results from a user study conducted to validate the interface for use as an general purpose SSD in addition to use in interactive applications. The chapter concludes with a discussion of the limitations of the device and the need for a more involved mechanic for active exploration in order for the SSD to be useful in richer (and more realistic) environments. Chapter 6 introduces this mechanic, Foveated Haptic Gaze (FHG), a technique which takes inspiration from the foveated nature of the human visual system that enables haptic SSDs to intuitively convey dynamic information in interactive domains. The chapter also discusses the results of a user study, which included both sighted and individuals who are blind as participants, conducted to assess the efficacy of FHG in an interactive 3D game based on DOOM, a classic first-person shooter game. Finally, the chapter discusses the future direction of this new branch of interactive SSDs and the barriers that still exist for the technology

to be useful in real-world scenarios. Chapter 7 introduces a novel, person-centered approach to one-shot object detection called 1SODDA as a step towards generalizing these Sensory Substitution techniques to real-world scenarios. The method leverages Deep Object Detection techniques and a person-centered data collection and augmentation approach that places the user in control by allowing them to specify objects of interest for the model to detect. Lastly, Chapter 8 concludes the dissertation with discussion of the findings and the future direction of modern Sensory Substitution in tandem with Computer Vision.

Chapter 2

BACKGROUND

In order to develop assistive technology tasked with communicating vision via haptics in a principled manner, an understanding of the underlying biological and cognitive processes for these modalities is necessary. These processes are explored in this chapter with a focus on the sub-topics most pertinant to the development of haptic Sensory Substitution Devices for vision. This chapter also includes an in-depth discussion on Sensory Substitution, its applications, limitations, and promising directions.

2.1 The Human Visual System

The Human Visual System (HSV) consists of three primary components: they eye, optic nerve, and visual cortex. The eye focuses incoming light onto the retina which converts the electromagnetic energy into electrical impulses. These impulses are carried by the optic nerve to the visual cortex in the brain, where the signal is processed. Figure 2.1 illustrates the anatomy of the human eye: incoming light is refracted by both the cornea and lens to create a focused image onto the retina. The optical axis, or center of the path by which light travels through the eye, is aligned with the fovea: a portion of the retina corresponding to the center of the field of view.

The retina contains photosensitive cells called photoceptors, which upon absorbing light generate a neural signal. Humans possess two kinds of photoreceptors: rods and cones. Rods are highly sensitive cells that can respond to as few photons as one, while cones require much more light and are responsible for color vision. There are 3 types of cones roughly corresponding to the wavelengths they are sensitive to: long-wave

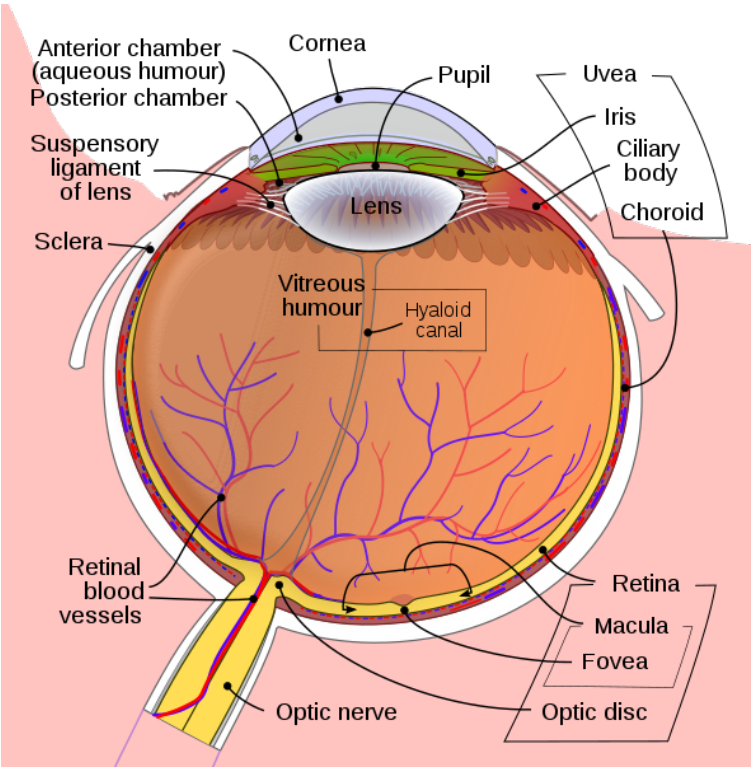


Figure 2.1: Anatomical diagram of the human eye. *Adapted from Wikimedia commons: Schematic_diagram_of_the_human_eye_en.svg,* ***https://en.wikipedia.org/wiki/File:Schematic_diagram_of_the_human_eye_en.svg***

(L), medium-wave (M), and short-wave (S). Their sensitivities are plotted against wavelength in figure 2.2

The fovea has the highest density of photoreceptors and is populated exclusively by cones. In addition to the increased density at the fovea, the foveala, a portion at the center of the fovea, the cells that lay over the photoreceptors in the rest of the retina are aranged in a manner that prevents obstruction of the light before reaching the photoreceptors. The increased density of photoreceptors in tandem with their unobstructed line-of-sight make the foveola the most acute portion of the retina (illustrated in figure 2.4). The fovea encompasses a mere 1% of the retinal area and 2 deg of the visual field, while corresponding to 50% of the visual cortex,

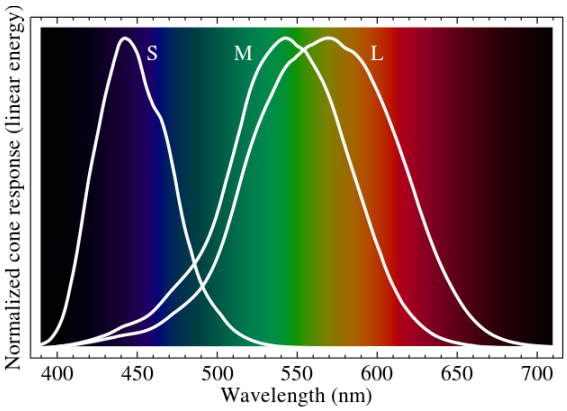


Figure 2.2: Normalized cone sensitivity over the entire visible spectrum. *Adapted from Wikimedia commons: Cone-fundamentals-with-srgb-spectrum.svg,* ***https://en.wikipedia.org/wiki/File:Cone-fundamentals-with-srgb-spectrum.svg***

Krantz (2012), Zhu and Yang (2002). This disproportionate allotment illustrates the importance of this narrow portion of human vision. Figure 2.1 illustrates the relative color and spatial acuity in the foveated region versus the rest of the field of view. After signal travels through a collection of low-level processing circuitry (bipolar cells and retinal ganglion cells) and finally arrives at the optic nerve where it is sent to the optical cortex. For a more in-depth description of the HSV, see Hudspeth, A.J.; Schwartz, James; Siegelbaum, Steven; Kandel, Eric; Jessell (2012).

In addition to progressive resolution of the human visual system, the attentional mechanisms in human visual processing are also noteworthy. The work by Neisser, Neisser and Becklen (1975), which was later expounded upon by Simons and Chabris, Simons and Chabris (1999), on inattentional blindness explored the phenomena that humans are often unaware of objects in their field of view that are outside of their attentional focus. In possibly the most famous of these studies, Simons and Chabris tasked participants to count the number of passes two teams made with a basketball in a pre-recorded video. While most of the participants had little trouble count-

(a) (b)

Figure 2.3: (a) Original image of flower (b) Notice the stark difference in visual acuity in both color and sharpness inside versus outside of the foveated region.

ing the passes accurately, a surprising number (27-58% of participants, depending on conditions) failed to notice an anamolis event event occuring during the video (a woman walking through with an umbrella, and even more surprisingly a person walking through in a full gorilla suit). The phenomena has been reproduced in many other scenarios; participants in one study even failed to notice an ongoing street fight during one study, Chabris *et al.* (2011).

While initially unintuitive, the alternative, that humans perceive and process the entirety of their visual field simultaneously is even more doubtful. Perceptual bandwidth limits have been theorized for many of the modalities (elaborated upon in section 2.4) in addition to attentional limits. Attention therefore plays an immense role in visual processing. Simons and Chabris go so far as to assert that "conscious

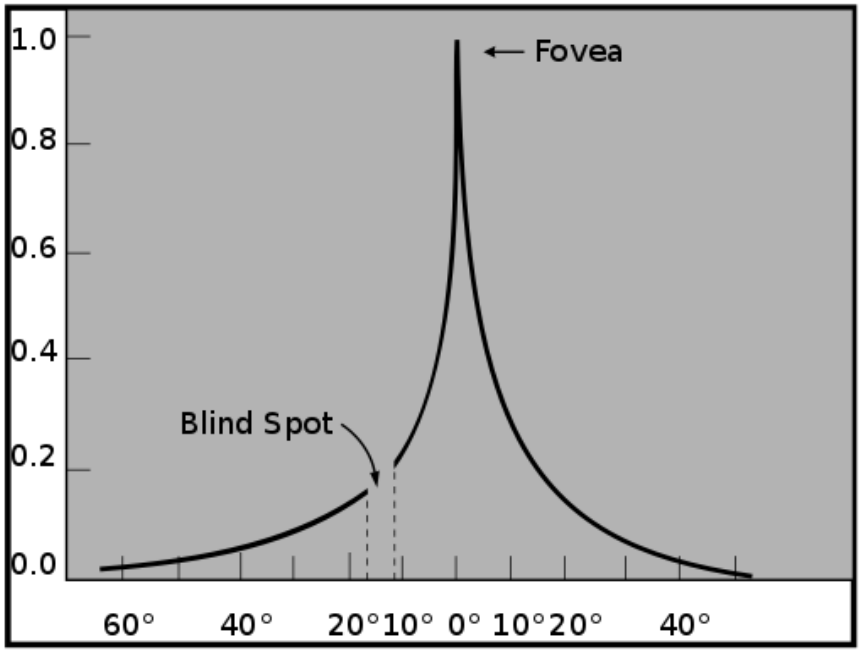


Figure 2.4: Visual Acuity vs Angle from Fovea Centralis. *Adapted from Wikimedia commons: AcuityHumanEye.svg,* `https://commons.wikimedia.org/wiki/File:AcuityHumanEye.svg`

perception seems to require attention", Simons and Chabris (1999), and make the observation that when the unexpected event contains visual characteristics similar to the attended event the participant was more likely to notice the unexpected event. This observation implies that when primed with a specific task, visual characteristics that correspond to that task are more salient than those outside of the attended tasks scope.

2.2 Haptics and Perception

The basis for the perception of haptic stimuli are mechanoreceptors embedded in the skin. Mechanoreceptors are biological transducers that convert mechanical stimulation into electrical signals (action/graded potentials) to be transmitted by

nerves. The human skin contains several kinds: *Ruffini Endings* which respond to skin stretching, *Merkel Cells* which respond to low-frequency stimuli such as points and edges of objects, *Pacinian Corpuscles* which respond to vibrations, and *Meissner Corpuscles* which detect lateral skin motion, *Field Receptors* which respond to slow stroking stimuli, and hair follicle receptors which detect hair movement, Hudspeth, A.J.; Schwartz, James; Siegelbaum, Steven; Kandel, Eric; Jessell (2012). The properties of mechanoreceptors must be considered when designing a Haptic SSD.

Pacinian Corpuscles are of specific interest, as they respond to vibration in the range of 5-1000Hz, and are most sensitive to vibration at 200Hz. It is no coincidence that the pancake motors used to generate vibration in user devices vibrate close to this range. The motors used in the LRHI, Fakhri *et al.* (2019), for example vibrate at 220Hz, Type and Features (2016), very close to peak sensitivity. Pacinian Corpuscles are also the most sensitive with respect to indentation depth, capable of detecting skin indentations as small as 0.01 μm, which is magnitudes smaller than the second closest mechanoreceptor the Messner Corpuscle at 2 μm. The Pacinian Corpuscle is also a "Rapidly Adapting" mechanoreceptor, meaning it stops firing shortly after a stimulus stops changing, similar to retinal cells in the human visual system, Baccus and Meister (2002). Both of these properties are advantages of the Pacinian Corpuscle as a target for Haptic SSDs as inexpensive and low profile vibration motors are relatively easy integrate into an SSD. A drawback of the Pacinian Corpuscle though is a receptive field that is larger than most other mechanoreceptors, as they lay deeper in the skin (2-3mm deep) and are more sparsely located. Practically this means localization of vibrations is less precise than that of other stimuli. While the two-point discrimination threshold of vibrotactile stimulation on a human back is in the range of 10-11mm, Jones and Sarter (2008), localizing specific stimuli in an array

of actuators tends to be less precise. Researchers found that using a 3x3 vibrotactile array on the back with intermotor spacing of 6cm, participants were able to localize the vibrations at an accuracy of 84%, Lindeman and Yanagida (2003). A more in-depth study on the perception of vibrotacile arrays on the human back was done by Jones et al. who developed a 4x4 array with a vertical spacing of 4cm and horizontal spacing of 6cm. With this higher density array, localization accuracy of individual motors dropped to 59%, while disciminating haptic patterns was a much easier task. Participants achieved an accuracy of 95% for haptic pattern recognition, Jones and Ray (2008). Consequently, for haptic arrays to be placed on the back, the limit for accurate actuator localization appears to exist between 4-6cm spacing, while the limit for vibrotactile pattern discrimination remains higher.

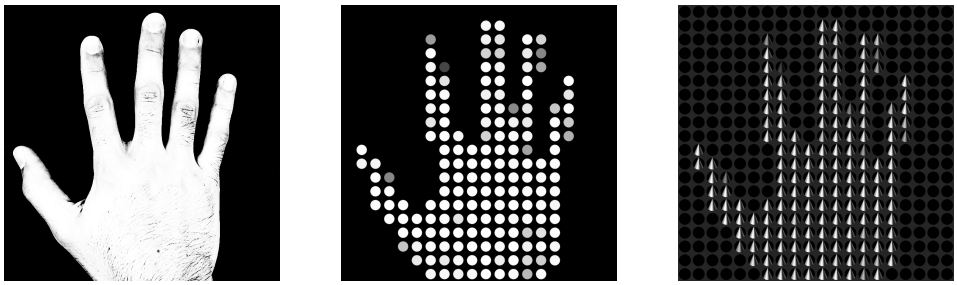
Haptic arrays relying on static pressure versus vibration can get away with denser spacing. The TVSS for example made use of solenoids spaced only 12mm apart, resulting in an array with 400 total actuators, Bach-Y-Rita *et al.* (1969). This is possible due to the smaller receptive fields of the mechanoreceptors targeted by the solenoids (Merkel Cells). While higher density haptic arrays provide more higher spatial resolution, the cost, size, and power requirements of solenoids often make them impractical for Sensory Substitution applications in comparison to vibration motors. Lower resolution displays may also be completely adequate for applications conveying information that is general and relative versus specific and absolute, as those discussed in section 6.

2.3 Sensory Substitution

Sensory Substitution (SS) is the process of delivering a signal from the domain of one sensory modality to an alternative sensory modality, for example circumventing

the auditory modality with the haptic modality. The purpose of SS is often to circumvent an impaired modality via an alternative one so that a person can experience stimuli from the impaired modality. Formally, we will refer to the modality being replaced as the *source modality* and the modality that the signal is being delivered to the *target modality*. In other words, Sensory Substitution is a method by which people who are blind can see by hearing, or people who are deaf can hear via touch. This revolutionary idea, that people can learn to experience sensations grounded in one modality via another, was pioneered by the late Dr. Bach-y-Rita in the 1960s. The notion that the signals eminating from the receptors of one modality can be interpreted in the brain as stimuli from another domain was novel and spurred the development of methods and systems in the last four decades harnessing this phenomenon to treat disability, enhance education, and enrich people's lives.

The objective of a Sensory Substitution Device (SSD) is to transform a signal from the source domain into a form that can be perceived by the target modality. In the famous case of Dr. Bach-y-Rita's TVSS system, the source domain was visual and the target domain haptics, Bach-Y-Rita *et al.* (1969). Dr. Bach-y-Rita showed that people who are blind could, with training, learn to interpret visual stimuli projected on their back as tactile stimuli using the Tactile Vision Sensory Substitution device (TVSS). The device consisted of a dental chair, retro-fitted with 400 solenoid actuators that would press upon the user's back when seated. The solenoids were controlled by a camera system that converted images to electrical signals: a bright portion of an image would result in solenoids pressing against the back of the user in the corresponding location. This is illustrated in fig. 2.5a, 2.5b, and 2.5c.



(a) (b) (c)

Figure 2.5: (a) Original image of hand (b) Original image converted into electrical activations based on the brightness of that portion of the image (c) activations are converted into solenoid positions to stimulate the skin: solenoids stimulate the skin in proportion to their activation.

After training, the TVSS allowed user who are blind to recognize household objects without touching them. These results had massive implications for the field of neuroscience and that of assistive technology - technology to improve the lives of people with disabilities, demonstrating that through clever uses of technology sensory impairments can be circumvented. Researchers later went on to develop Sensory Substitution Devices (SSDs) to substitute vision with hearing, Meijer (1992), vestibular with tactile, Bach-y Rita *et al.* (2005), and hearing for tactile, Novich (2015) with impressive initial results. While fantastic medical advances in sensory prosthesis such as the Cochlear Implant (CI), Merzenich *et al.* (1973) and retinal prosthesis, Caspi *et al.* (2009); SSDs provide a great alternative for circumventing the loss of a sensory modality as surgical procedures are often prohibitively expensive and always invasive. This dissertation will focus on SSDs with a source modality of vision and a target modality of haptics, or Haptic SSDs for Vision.

2.4 Advantages and Limitations of Haptics for Sensory Substitution

The skin is the largest organ on the body, making touch one of the most versatile modalities to design SSDs for. Designers have a wide range of options with respect to where to place devices: some SSDs have even been designed for the tongue. Because of the plentiful real-estate, haptic SSDs can be designed to impart minimal obstruction to other crucial functions of the senses. For example, haptic actuators can be placed on places such as the back, upper arms, or waistline, locations that are not often used in day-to-day activity.

Touch also happens to be underutilized as a communication medium for technology. Designing touch-based SSDs has the added benefit of likely not interfering with other communication mediums to cause sensory overload. A person who is blind for example is unlikely to accept obstructing their hearing with a vision-to-auditory SSD, but is more likely to if the target modality is one that is not already being highly utilized, such as touch. Using haptics not only avoids interfering with a modality already in use, but may allow for a higher effective cognitive bandwidth due to the multi-channel nature of adding haptics.

Current models of the human memory and attention system portray the different modalities as semi-independent channels to one's attention. The Baddeley multi-channel model for example allocates different sensory inputs unique and semi-independent subsystems of working memory and independent processing systems for such each modality, Sweller *et al.* (2011). Consequently, sensory signals of different modalities can more effectively make use of the human cognitive bandwidth than the same information presented to a single modality; this phenomena is called the

"modality effect". Sensory overload occurs when the attention system is overwhelmed and because touch is often underutilized in daily tasks, taking advantage of it can augment attentional bandwidth while successfully averting sensory overload. For this reason, the haptic modality has received substantial interest in military (high cognitive load) settings and haptic-vestibular SSDs have been developed for pilots flying in low-visibility settings, Van Erp and Self (2008).

The sense of touch though does exhibit inherent limitations. One such limitation is the limited information capacity of haptics. It is estimated that the visual system has a capacity of 4.3 Mbits/second, Jacobson (1951) and the auditory system 8 Kbits/second, Homer Jacobson (1950). In comparison, the haptic modality is estimated to have a mere 600-925 bits/second of capacity, Novich and Eagleman (2015). This implies an upper bound on the amount of information an SSD can convey through the sense of touch, and consequently an upper bound on the fidelity of information one can access from a higher-bandwidth modality through haptics. It was not so minuscule though that users could not use use it to substitute vision and perform basic vision tasks, Bach-Y-Rita *et al.* (1969).

While touch allows for a wide variety of locations, sophisticated interaction often requires multiple tactile actuators to convey complex information. The skin imposes a minimum spacing requirement between tactors to maintain discernability and this spacing is a function of the location of the body the tactors are placed as well as the kind of stimulation (pressure, vibration, temperature, etc) that will be applied. For example, on the human back the minimum discriminable separation of vibration stimuli is about 11mm, Eskildsen *et al.* (1969). Consequently, actuators must often be adequately spaced out on the body, taking up more space than a device relying

on a more concentrated modality like vision or hearing. The design of the device and the signal processing is crucial for effective use and adoption as an SSD and haptic SSDs can largely be categorized into 3 categories: general purpose (section 2.5), media readers (section 2.6), and interactive devices (section 3) which are explored in the following sections.

2.5 General Purpose Sensory Substitution

General purpose Sensory Substitution is intended to circumvent a source modality via the target modality outright, making it a complete substitute for the source modality. This is in contrast to application specific SS, where a device or technique transforms a signal from the source modality into the domain of the target modality in such a way that is tailored to the application. Oftentimes there exists a tradeoff between efficiency and generality: the more general an SS method, the more training is reqiured, while more application-specific methods are often learned more quickly.

The first and likely most famous implementation of general purpose vision sensory substitution occured in 1969, when Dr. Paul Bach-y-Rita and his team developed the Tactile Vision Sensory Substitution (TVSS), demonstrating that with a somewhat long training period (up to 150 hours), users of the device could recognize common objects as well as motion, gradients, and shadows at a distance, Bach-Y-Rita *et al.* (1969) and White *et al.* (1970). While impressive, the work had a long way to go towards vision-to-tactile SS that could truly replace vision, let alone be a practical solution for daily activity. The device was incredibly bulky, having been constructed from a dental chair, hundreds of solenoids, camera equipment and electrical amplifiers. The device's resolution was also too low to discern fine detail and long training

times were required for proficiency. Furthermore, the system lacked color detection and sported a field of view was that was narrow and fixed. All of these problems made it impractical for real-world use such as navigation, reading, etc.

Some of these issues were addressed in later devices. For example the "Rabbit Display", developed by the MIT Media Lab, made use of a tactile illusion called "saltation" in order to increase the effective resolution of a low resolution tactile display, Tan and Pentland (1997). Saltation (also known as the "cutaneous rabbit") is an illusory sensation of touch felt in between the location of where the stimuli was actually applied to the skin, Geldard and Sherrick (1972) and can be achieved by timing the stimuli in a specific manner. The authors emphasized that the display would be useful in conveying direction information to users such as pilots (such as a vestibular SSD) or to help people with navigation. Because of the low resolution nature of the display (3x3), it can be inferred that it can be made relatively small and lightweight, making it a viable option for mobile applications and more socially acceptable. Saltation can even evoke senation away from the body, Miyazaki *et al.* (2010), and may be used in the future to "extend" displays off of the body. The low-resolution nature of the display though limits the detail that can be conveyed, even if saltation is employed to increase perceived resolution. Generalizing this technique to a larger, finer display is not trivial though, as inducing saltation requires haptic stimuli to be presented to the skin in specific timings and patterns, limiting the representable patterns of the display and thus the informational content.

Further improving on acuity and portability, researchers in 2001 developed a Tongue Display Unit (TDU) for vision-to-electrotactile Sensory Substitution applications. The device converts images from a digital camera into electrical signals that

are applied to the tongue in a similar manner to how Bach-y-Rita's TVSS converted image information into tactile stimulation (illustrated in fig. 2.5c). While unconventional, the tongue was chosen as the site for the TDU for both its sensitivity to electric current and density of receptors, making it better suited for discerning fine details than a user's back. Researchers showed that users of the Tongue Display Unit were able to achieve a visual acuity of 20/860 on a standard "Tumbling E" visual acuity test and 20/430 after 9 hours of training, generalizing much better than the original TVSS, Sampaio *et al.* (2001) and Nau *et al.* (2013).

The same group went on to use the TDU as a rehabilitation device for people with vestibular conditions affecting their balance, renaming the TDU the BrainPort. Researchers used the BrainPort to convey balance information to people who had lost their sense of balance, substituting it with electrotactile stimulation and saw marked improvements in balance, some users being able to stop using the device entirely while retaining their newfound balance, Bach-y Rita *et al.* (2005). This group demonstrated that haptic SSDs can not only be used as sensory substitutes but also as rehabilitation devices.

With all of these advances since the original TVSS, there are some limitations that remain untouched such as color distinction and stereo vision. Vision to tactile SSDs also are still cumbersome for practical daily use as the state-of-the-art implementations (BrainPort) require the display to be in the mouth limiting social interactions and possibly exacerbating stigma towards users. There has been more success in general vision substitution with the auditory system as the target modality. Blind users have even been able to navigate with SSDs such as the "vOICe", which stands for "oh I see", Meijer (1992) and Ward and Meijer (2010), and experi-

ence color with EyeMusic, Abboud *et al.* (2014), a system that abstracts images into tones and sounds of instruments hence the name. The discrepancy in performance between vision-to-haptic and vision-to-auditory SSDs is likely to do the information capacity discrepancy being smaller between vision and auditory versus vision and haptics. Auditory-to-Haptic Sensory Substitution though enjoys a similar advantage over vision-to-haptic.

Some of the earliest attempts at general auditory-to-haptic SS were made by the Audiological Engineering Corp in the 1980s. The group designed what are now known as the Tactaid devices. The devices partition audio data into a varying number of bands based on the model of Tactaid device; for example, Tactaid VII uses seven bands and conveys activity in the bands to the user via seven unique vibrotactile actuators. Researchers evaluated the devices with users who had hearing impairments and found that users were able to discern syllables and showed "enhanced monosyllabic word recognition" but users did not report significant subjective improvements in recognition of speech, Karyn *et al.* (1999). A more recent and more successful method for auditory-to-haptic SS was developed in 2014 by researchers at Rice University. Instead of using just seven tactors, researchers developed a suit called the VEST containing 26 eccentric rotating mass (ERM) motors, developing patterns involving groups of 9 vibrotactile motors in a square array that conveyed directional "sweeps". They found that the spatiotemporal sweep patterns were more distinguishable than just spatial or static patterns alone. Combining the VEST with speech processing methods (compressing and converting the speech into haptic patterns), users were able to discern speech much more clearly than ever before, distinguishing words at much higher accuracies than with the Tactaid devices, Eagleman (2014) and Novich (2015). General purpose Sensory Substitution devices explore the limits of percep-

tion but are rarely ever widely adopted as assistive technology. Instead, application specific SSDs tend to have more success as practical aids for daily use.

2.6 Media

While the written word enabled mass communication, standard media formats are not accessible to the entirety of society. People with visual impairments often have difficulty accessing communication mediums due to their design being reliant on vision. Haptic SSDs for reading media are designed to convey the information in media that is visual or text-based to the sense of touch. Examples of this include devices for reading text, exploring images, and understanding maps, which are all important for and individual's education and independence.

2.7 Language and Communication

The most famous, and arguably most successful Sensory Substitution technique for reading is Braille, the tactile, two column, three row cells of raised dots were invented by Louis Braille and published in 1829. Alphanumeric characters are converted into tactile representations, where each letter or number is assigned a Braille code occupying one Braille cell. These cells can be read and written and are the standard reading and writing system for individuals who are blind in many countries. It has been shown that after extenstive practice Braille users can achieve a reading rate of 90wpm, Troxel (1967). Visual reading rates are about 200 words per minute for comparison. Written text though had to be translated into Braille before it was accessible and was often bulkier than the original material. Almost a decade and a half after the invention of Braille, refreshable Braille displays emerged as a solution

to the size and heft of translated works. Refreshable Braille displays typically consist of a row of refreshable Braille cells where the dots are controlled by a piezoelectric bimorph cantilever that is activated by an electric potential, Smithmaitrie *et al.* (2008). Modern refreshable Braille cells are 2x4 in contrast to the original 2x3 cells, with the additional 2 dots used for cursor position and other indicators, according to the American Foundation for the Blind (AFB), Stageberg (2004). Modern refreshable displays sport between 18 and 84 Braille cells and can interface with computers via bluetooth, while also utilizing input controls for typing and navigating, Schmidt *et al.* (1998), Bucchieri (2013) and Freedom Scientific Inc. (2018). In response to the era of touchscreens, methods for typing in Braille have been evolved to be compatible with the flat featureless surfaces of touchscreens, Mascetti *et al.* (2011). Many touchscreen consumer devices today allow for Braille typing using solely the display, eliminating the requirement for additional hardware.

Other less successful methods for tactile communication were developed such as Vibratese, a tactile language based on vibrations on the body that varied in amplitude and duration to communicate alphanumeric symbols. Invented by F. A. Geldard in 1957, test subjects trained in Vibratese were able to achieve a reading rate of up to 60 words per minute using the system, Pasquero (2006). The language never saw widespread adoption likely due to Braille having already been the standard.

Another alternative to the refreshable Braille cell called *STRESS* emerged in 2005. The device uses vertical stacks of piezoelectric plates that deform with an electric current. The user places their fingers on top of the stack so that their fingers are perpendicular to the individual piezoelectric plates and the plates bend in response to applied electric current to create different sensations at the finger tips. Researchers

saw promising preliminary results in creating "virtual Braille" with the *STRESS* device, a 1-dimensional version of Braille, Lévesque *et al.* (2005). Researchers then went on to explore more complicated game based use-cases with the technology, Wang *et al.* (2006), detailed in section 3.4.

Unfortunately, while technological improvements continue to advance the defacto tactile communication method Braille, literacy is in decline. The National Federation of the Blind (NFB) in a 2009 report stated that the Braille literacy rate has dwindled to less than 10% of individuals who are blind in the United States, Jernigan Institute (2009). The NFB report states that Braille education is critical to literacy and employment among individuals who are blind, and while screen-readers have facilitated computer access their existence likely inhibits Braille adaption. The NFB is calling for Braille adoption to be elevated in priority for those who teach individuals who are blind.

One of the issues with Braille though is that non-digital text must be transcribed before becoming accessible, and in response several technologies emerged to read written characters beginning in the 1960s. The Optohapt for example used photosensitive sensors to detect characters on paper (on a retrofitted typewriter). The characters were passed through the sensor at a rate of 70 characters per minute creating electrical signals that were sent to vibrating actuators located at 9 spatially dispered bodily sites, Geldard (1966). During the same period, a competing device proposed by Linvill and Bliss called the Optacon (OPtical-to-TActile CONversion), was developed. The device consisted of a capture module (a wand-like device) fitted with an 8x12 array of photosensitive cells can be placed on a page to be read with a user's dominant hand. The user then places a finger from their other hand on the actua-

(a) (b) (c)

Figure 2.6: (a) Original image representation of letter **F** (b) Letter converted into electrical activations (c) activations converted into solenoid positions to stimulate the skin.

tor. The actuator is an array of 24x6 pins that the finger rests on that move up and down in response to signals from the capture module, Bliss *et al.* (1970). The authors claimed that a reading rate of 50 wpm could be achived with 160 hours of training.

The researchers behind the TVSS also explored different ways to display letters using the device instead of visual information, comparing static haptic patterns and dynamic ones for each letter. Letters were converted to tactile stimuli (illustrated in fig. 2.6a, 2.6b, 2.6c). They found that a sliding window approach was most successful for accurate letter discimination among participants in a user study, achieving an accuracy of 51% correct letter discrimination, Loomis (1974). This conclusion (that spatiotemporal patterns are more discriminable than static ones) has been supported by later work by the developers of the VEST, Novich (2015) and LRHI, Fakhri *et al.* (2019). The sliding window approach only exposed a user to a portion of the letter at any one moment, but the whole letter would be presented over a duration of 1 second, illustrated in fig. 2.7, imposing a maximum reading rate of 60 characters per minute (60cpm).

(a) (b) (c)

Figure 2.7: Sliding window presentation of the letter **F**. A user would be exposed to the sliding window stimuli over about 1 second.

The TVSS implementation of a character reader seemed like overkill (400 actuators), and with a limit of 60cpm it did not show much promise as a media reading device. More recently lower resolution displays have been explored for communicating written characters. Researchers developed a low resolution tactor array of 9 vibration actuators placed 3x3 on the back rest of a chair. Representations of letters were "traced" over the tactors as if the letters were being dynamically "drawn" on the user's back. Patterns varried over space and time and participants were able to achieve an accuracy of 87% for letter and number recognition, Yanagida *et al.* (2004). This was vastly higher than the accuracies achieved with the TVSS (51%) with far fewer actuators. This leads us to believe that for abstract information representation a more "coded" scheme may be more useful than attempting to reproduce the characteristics of the visual content faithfully. Although the hardware requirements are vastly reduced and accuracies improved, the dynamic patterns may still be too slow for use in real time, implying that a different coding scheme similar to Braille may be more practically useful.

Braille-like devices are still superior it seems when it comes to reading and writing using haptics and while Braille may be currently in decline, emerging technology in

the space of refreshable Braille displays has appeared as recently as 2017 in the form of non-mechanical, air actuated displays in contrast to piezoelectric designs. This new technology uses fluids to make bubbles in the display as the dots, and it is being integrated with a traditional touchscreen tablet. This technology appeared in 2017 in the form of the Blitab (a play on words combining "blind" and "tablet") and is purported to have 14 rows of 23 6-dot Braille cells, Metz (2017). This 2-dimensional display paves the way for richer human-computer-interaction and possibly a reemergence of Braille literacy.

2.8 Visual Content Readers

Apart from language systems, there has been growing interest in the development of haptic devices for understanding traditionally visual information such as images, graphs, and maps. Students with visual impairments are often at a disadvantage in academic settings because the content is in an inaccessible format. Even when text is transcribed or conveyed via a media reading device, images continue to present a challenge to students and teachers. An intuitive method for representing two-dimensional information using haptics are "raised paper diagrams". These diagrams are often made from "swell paper", which expands in an oven-like device where it has been printed on creating a tactile surface, Miller *et al.* (2011). An example of such a diagram is illustrated in fig. 2.8a and 2.8b. A similar method for creating 2D tactile visualizations that allows an end-user to reconfigure a diagram is in the form of moldable wax-based rods called "Wikki Stix", shown in fig. 2.9. Users can scan them with their fingers to feel the features of the visualization. While useful, Wikki Stix and raised paper diagrams still requires a translation from an original image for instructional purposes. It is also often difficult to incorporate sufficient information

(a) (b)

Figure 2.8: (a) Raised paper diagram of a man's head on white paper (b) The same diagram viewed close and at an angle. Note that several different heights and thicknesses are possible on such diagrams.

density due to the physical limitations of the media. Descriptions are often added by a teacher or caption to aid in comprehension of the visualizations, but a more elegant solution has been developed in the form of the Talking Tactile Tablet (T3). The T3 consists of a tactile diagram that can be felt overlaying a touch sensitive screen. When a user presses the tactile map they are presented with auxiliary audible information to complement the tactile map, Landau and Wells (2003). An even more fleshed out version uses a smartphone and 3D printed overlays to perform a similarly multimodal experience to the T3, is called TacTILE. Authors of the TacTILE developed a complete toolchain for the rapid development of such devices, He *et al.* (2017).

More elaborate attempts to make visual information accessible began appearing in the late 1990s. Japanese researchers Ikei et al. attempted to convey an image's textures via haptics by constructing a 5x10 pin finger display driven by piezoelectric actuators (similar to refreshable Braille displays). The pins though were not static like their Braille counterparts, but vibrated at 250Hz at varying amplitudes to mimic tactile textures. Researchers converted close-up images of textured surfaces such as

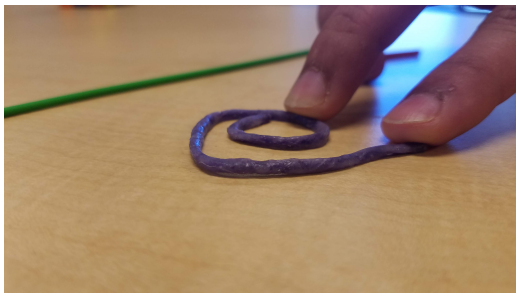


Figure 2.9: Wikki Stix used for conveying visual-spatial information via haptics. They are flexible and waxy, making them easily configurable and stationary on surfaces.

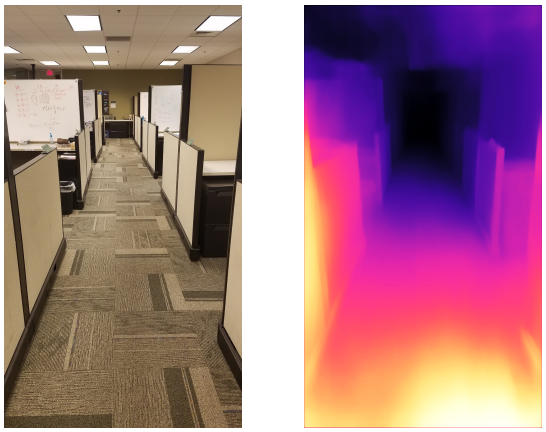
a bamboo woven basket, thatch basket, painted wall, and a rug to haptic textures by converting the images to pin intensities on their finger display. A user study revealed that using their technique sighted users were able to correctly identify the image belonging to the texture being displayed on the finger pad more than 90% of the time, Ikei *et al.* (1997). The high recognition accuracies and straight-forward method for converting images to tactile representations was promising, as generalizing to other domains would be relatively simple, although no study was performed with individuals who are blind and thus had no visual reference for the textures they were experiencing. Ikei's method worked for any arbitrary texture but had no sense of "space" that is required to accurately convey most visualizations.

Researchers Wall and Brewster sought to solve this problem in 2006 when they developed a graphical diagram reading system by integrating the VTPlayer mouse with a digital drawing tablet and used the stylus to interact with the graph. The VTPlayer mouse is a computer mouse that is augmented with two 4x4-pin Braille cells. The user would point on the tablet with the stylus and receive textured information of what they were pointing at with the VTPlayer on their non-dominant hand. Complementary audio feedback would also be available if the user pressed the buttons on

the VTPlayer, Wall and Brewster (2006a). Earlier, in 2005, Wall and Brewster performed a psychophysical study comparing the TVPlayer mouse, the WingMan Force Feedback mouse, and classic raised paper for use in image understanding. They used a simple line gradient discrimination task: a line was displayed and participants were asked to disciminate the gradient of the line using the three devices. While the force feedback mouse outperformed the VTPlayer, the raised paper was superior. Interestingly, the authors surmised that this is likely due to the combination of proprioceptive and tactile cues that neither the VTPlayer or WingMan mouse provide at the same time, Wall and Brewster (2006b), which likely led them to develop the 2006 graph reading system using a stylus as well as the tactile feedback from the VTPlayer mouse.

2.9 Future Trends for Sensory Substitution

Some standout implementations of Haptic Sensory Substitution are Bach-y-Rita's TVSS, the BrainPort, and the Eagleman and Novich's VEST, showing the true raw representational power of the modality, but they also reveal some limitations. For the TVSS, long training hours, a chair-based design with many actuators, and lack of fine details hinder its use in real-world applications. While the BrainPort tackles the portability and details issues somewhat, it still suffers from the practical concerns of requiring the display to be placed on a user's tongue. For auditory substitution, the VEST is impressive in its ability to convey speech, but other more subtle aspects of hearing are still missing, such as localization via stereo hearing. Further strides in the realm of Haptic Sensory Substitution are more likely to arise with clever integrations with emerging signal processing tools and clever delivery techniques.



(a) (b)

Figure 2.10: (a) Original image of an office (b) depth image from model trained on the MegaDepth depth dataset, Li and Snavely (2018).

In the realm of vision-to-haptic SS, strides in Computer Vision show promise for enabling more effective Sensory Substitution. For example, object detection has made great strides, as well as depth estimation from monocular images. Having access to both depth and object identities from monocular images could drastically improve ETAs by allowing ones that rely on depth information to use only a camera instead of lasers, sonar, infrared, or stereo cameras. Fig. 2.10b illustrates the impressive performance of emerging depth estimation models (MegaDepth). The methods underlying the image understanding applications from section 2.8 utilizing neural networks also show great promise in augmenting haptic SSD technology. Combining these powerful models with a proprioceptive and tactile interace would likely lead to a more effective and meaningful image understanding tool that can be used both in the physical world but even more so in virtual environments.

Chapter 3

RELATED WORKS

Sensory Substitution devices for interactive applications are designed to function in environments that respond or change with respect to the user's behavior. For example, playing a video game is interactive while reading a textbook is not. Thus SSDs for interactive applications must contend with the demands of interactive environments, that is latency sensitivity, sensory overload, and diverse and dynamic situations. This section will explore SSDs designed for interactive applications of mobility and travel, interactive instructional systems, social interactions, and virtual interactive environments.

3.1 Instructional Systems

Instructional SSDs are those that are intended to be used for learning; more specifically they are intended to be used for learning in dynamic environments that react to user input, in contrast to media reading SSDs that are intended to be used to convey information about static sources such as books and illustrations.

3.1.1 Mobility Learning

Sighted people can look up images of a location and quickly acquaint themselves with the flow of the environment. Unfortunately, those for who images are inaccessible do not have such a luxury and can not benefit from the vast amounts of visual data that is available online. Furthermore, familiarity with an environment is often

Figure 3.1: Sensable Inc's PHANToM Desktop, a force feedback device for haptic applications

more important for people with visual impairments than sighted individuals. To address this issue, virtual environments that model locations that are of interest and allow people with visual impairments to interact with those environments may benefit people with visual impairments by allowing them to familiarize themselves with the novel location before visiting in person. These systems are refered to as "Mobility Training" systems.

On such system developed at the University of Colorado at Colorado Springs is called MoVE: Mobiltiy Training in Haptic Virtual Environment Semwal (2001). Its purpose is to enable people who are blind to explore a model of new environments haptically. The system is iterative, a user explores the virtual model, then explores the physical location and repeats this process to fine-tune their understanding of the space, intuitively learning the relationship between the rendered world and the real world. MoVE uses SensAble Inc's PHANToM force feedback device, allowing users to

interact with the virtual environment by poking around with the PHANToM (shown in fig. 3.1), receiving force feedback when they contact objects. In a preliminary study, researchers found that user who are blind were quickly able to discriminate simple virtual objects such as spheres versus planes. While this approach is promising, the iterative nature has yet to be tested for individuals with visual impairments.

Sharkey et al. devised a more comprehensive approach using a force feedback joystick, audio feedback, and a "guiding computer agent" to create and explore virtual environments before exploring their real counterparts they were modelled after. The force feedback encoded information about texture, objects via force-fields, and structural boundaries while the audio component added descriptions of the scene as well as of the user's orientation in space to aid in navigation. They found that users were able to accurately and quickly learn to navigate in the virtual environment and when presented with the physical version quickly generalized what they had learned to the real environment Sharkey *et al.* (2002). Later came Omero, combining haptic and acoustic feedback with user preferences to learn the layout of new locations similar to the Skarkey system. Researchers tested the system with people with visual impairments and received positive subjective feedback; those with really low vision were not as successful as the system made extended use of visualizations on a monitor De Felice *et al.* (2007).

Lahav et al. developed a similar system for cognitive mapping via a multimodal approach and compared the performance of users who are blind in real-world navigation tasks versus other users who did not have access to the technology, expressing that users who had access to the technology developed more complete and accurate cognitive maps of the environment Lahav and Mioduser (2008). Researchers used

a multisensory virtual environment (MVE) that individuals who are blind could explore before exploring a physical environment (laid out in the same way). The MVE provided haptic force feedback and audio feedback of obstacles in the environment. Researchers found that individuals who are blind and were allowed to use the MVE developed more complete and accurate cognitive maps of the environment than those who were not given access to the MVE.

A more realistic approach was designed by Tzovaras et al. in 2009: a mixed reality system for training/educating people who are blind using a virtual white cane via the CyberGrasp device. Using a virtual white cane, trainees were able to traverse a life-sized virtual replica of an environment. Researchers enhanced the experience by providing realistic haptic feedback of cane collisions with virtual objects and realistic audio feedback Tzovaras *et al.* (2009). This method provided the most realistic approach as users employed skills to navigate the real environment almost identically to the virtual one but may not have been the most effective for generating complete cognitive maps of the environment. A direct comparison of this mixed reality real-scale method and the non-virtual reality methods above would be a welcome addition to the literature to unveil specific advantages and disadvantages of the two approaches. Furthermore, all of these Mobility Training systems require designers to model the environments beforehand, effectively reducing the pool of available environments to a small batch. This could possibly be rectified with crowdsourcing and integration with 2D to 3D modelling techniques.

3.1.2 Motor Learning

Motor learning is the development of motor skills, and motor learning tools are tools that aid in the development of such skills. In many motor learning settings demonstrations make up the majority of the instruction. Visual impairments can hinder this kind of instruction and haptic SSDs provide a valuable avenue to replace visual instruction. Motor learning systems may also provide feedback with respect to a user's movement in real-time, something that an instructor may not be able to give. Furthermore, some users may not be receptive to touch-based feedback from an instructor and may feel more comfortable with a device's feedback to correct motor movements. In the absence of an in-person instructor, or when an instructor does not have time to devote to a single student, an SSD that conveys motor skill information would be also be useful to most users.

In 2002, Yang et al. designed a suit for VR-based motor learning covering the torso with a vibrotactile display called POS.T. Wear. Employing a technique called "Just Follow Me" (JFM), the researchers used the POS.T. Wear to convey movement information of nearby objects to the wearer. The JFM metaphor consists of a "ghostly master" (illustrated in fig. 3.2) that is overlayed onto the trainee's body in the virtual environment. The master will then guide the trainee by performing the correct movements to be learned by the trainee. Yang et al. used JFM and the POS.T. Wear to study a user's obstacle awareness in virtual worlds and later as a motor learning tool Yang *et al.* (2002).

A more intuitive haptic motor learning approach called Mapping of Vibrations to Movement (MOVeMENT) was developed by McDaniel et al. Instead of the ghostly

master avatar approach in JFM, MOVeMENT seeks to map haptic stimulation to basic movements of the human body in an intuitive fashion. MOVeMENT is novel in that it is not application specific and can generalize to almost any motor learning activity. By targeting basic movements, MOVeMENT is capable of generalizing to almost any complex movement. Basic movements were developed by dividing the body via three planes that span three-dimensional space (sagittal, frontal, and horizontal planes). The planes ground the fundamental movements: extension or flexion is movement that increases or decreases respectively a joint angle in the sagittal plane, abduction or adduction refers to movement occuring in the frontal plane towards or away from the sagittal plane (respectively), and pronation or supination is rotation of a joint angle towards or away from the body from within the horizontal plane. McDaniel et al. designed haptic patterns to code for these five fundamental movements and used them as building blocks to describe more complex movements to a user using a push-pull metaphor to illicit movement in a certain plane. Participants in a preliminary study found the patterns intuitive and were able to discriminate them with high accuracy McDaniel *et al.* (2010).

3.2 Social Interaction

Social interaction is crucial to the well-being of individuals and this of course applies to people with disabilities. Unfortunately, many disabilities preclude individuals from equitable inclusion in all aspects of social activity. This can be due to practical issues or even socially constructed expectations of social interaction. Towards enriching the lives of people with disabilities by enabling a more equitable social experience, many researchers have sought to develop systems to rectify some of these inadequacies.

Figure 3.2: A visualization of the ghostly master metaphor. A trainee (solid) feels the ghost (transparent) as it moves through the trainee's body while performing an instructional movement. Original image from WikiHow (2019).

Researchers at Arizona State University for example have developed several SSD technologies for use in social situations. The "Haptic Belt" (shown in fig. 3.3) paired with a face detection system conveys the direction and distance of other people during a social interaction McDaniel *et al.* (2008). Tactile Rhythm was also explored in order to convey interpersonal distances to individuals who are blind McDaniel *et al.* (2009). These are coarse details of social interactions that are less accessible to people who are blind, but there are also very important fine details of social interaction that people who are blind miss out on too. An example of this would be facial expressions. At the same lab, researchers developed the "VibroGlove" a glove to convey facial expressions to people who are blind Krishna *et al.* (2010). A chair-based approach was also explored, showing promise of conveying facial expression information via "Facial Action Units", a system for describing facial expressions by their structural parts Bala *et al.* (2014). This culminated in a project called the Social Interaction

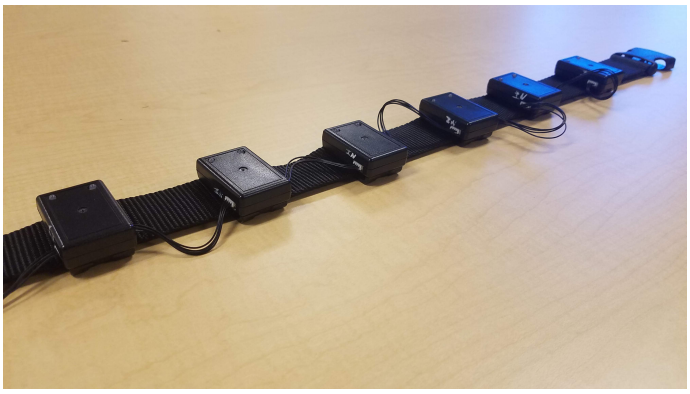


Figure 3.3: Haptic Belt developed at the CUbiC Lab at Arizona State University Rosenthal *et al.* (2011). The belt was designed to be modular and can be extended to fit more or fewer tactors connected in series. The location of the tactors can also be modified by simply sliding them along the belt.

Assistant (SIA), a person-centered SS system that combines active learning computer vision system with haptic tactors that convey information to users they might otherwise miss Panchanathan *et al.* (2016). A user would wear a camera similar to the shown in fig. 3.4a and receive haptic feedback from the camera using devices such as the VibroGlove and Haptic Belt (fig. 3.3).

3.3 Electronic Travel Aids (ETAs)

Mobility is a crucial component of independence, agency, and wellness. Vision disabilities account for a large portion of these mobility issues, and it is of no surprise because navigation itself is a complicated processes requiring visual integration over time and space and a strong dependence on memory. Researchers have determined that efficiently storing and recalling the relationship of landmarks in space is essential to spacial cognition, and thus navigation Monacelli *et al.* (2003), and because vision provides a method for establishing landmarks in 3D space it can be inferred that it

(a) (b)

Figure 3.4: (a) Mannequin wearing sunglasses mounted with a pinhole digital camera (b) close-up of pinhole camera

is heavily reliant on for navigation Ekstrom1 (2015). For this reason, a large number of SSDs have been developed to aid those with issues navigating. The most popular Sensory Substitution device for mobility is the "white cane", shown in fig. 3.5a and 3.5b. This device is used to transform information that would traditionally be acquired via vision to the haptic, priopreceptive, and auditory modalities. With the white cane, users scan the ground in front of them with the cane in sweeping motions in order to detect obstacles in their path by colliding with them. Users can often infer not just the existence of an obstacle but also some of the obstacle's properties via the tactile effects felt on contact as well as the sound emenating from the collision.

There are though drawbacks to the traditional "white cane" such as the limited range at which users can detect obstacles. White canes typically have a range of 1.5

(a) (b)

Figure 3.5: (a) PhD student Bryan Duarte navigating with white cane (b) close-up of white cane

meters in front of the user. A user must also collide with an object in order to detect it, which can be troublesome if the object is a person, dog, or something fragile. Users can also miss obstacles with the cane due to gaps in their sweeping pattern. White canes also can only detect obstacles at or below waist level, leaving the user vulnerable to obstacles like overhanging tree branches. Researchers have instinctually sought to improve upon the white cane to remedy some of these issues.

One of the first attempts to augment the white cane was in 1945 with the "Laser Cane". This device augmented a traditional cane with three gallium arsenide infrared laser rangefinders to detect obstacles and dropoffs at different distances. It was capable of detecting obstacles at several different angles, including an angle pointing upwards from the handle of the cane, so that users could detect obstacles above their

waist and avoid tree branches. Haptic and (optionally) audio feedback was delivered to the user based on the level and distance of a detected obstacle. The device was developed with continuous feedback from travelers who are blind and was finished in 1974 Benjamin (1974). While such a cane was novel, both laser and battery technology of the period restricted usage to a mere three hours per charge. The "Laser Cane" was one of the first attempts to give users information about obstacles before a collision, but it did so in a very coarse way, giving little information in the way of bearing (angle with respect to travel).

A method for detecting the bearing of objects was developed in 2002 by Dr. Roman Kuc. The device used two sonar range finders that together are used to infer the bearing of detected obstacles. Wrist-worn vibration motors vibrate with respect to the bearing of the obstacle, giving the user distance and direction information Kuc (2002). Several other "smart" canes were developed. Researchers at the Indian Institute of Technology performed a study, and found that their ultrasonic "Smart Cane" increased obstacle awareness, decreased collision prevelance, and increased mean detection distance as compared to traditional white canes in a navigation task Fallis (2010). Similar attempts at building smart canes are prevalent Menikdiwela *et al.* (2013) Ando *et al.* (2015) and are commonly variants of each other but GHARIEB and NAGIB (2015) takes the most elaborate approach whereby the cane is equipped with wheels and "drives" a user around. The device introduces modes such as "goal finding", where the device navigates for the user, providing turn by turn directions. This device though has not been verified by a study.

A more nuanced approach is the caneless ETA, removing altogether the need for a white cane. One such configuration is called the "Haptic Radar", a self-contained

headworn headband augmented with sensors that detects obstacles and intuitively conveys them to the wearer via haptics. The array of sensors each convey obstacle distance information for a path emanating from the sensor (there are several circling the head) Riener and Hartl (1974). Researchers found that participants tasked with a navigation task navigated more confidently with a Haptic Radar than without Cassinelli *et al.* (2014). Caneless systems may be advantages as they may reduce stigma induced by the iconic white cane. With GPS becoming ubiquitous, turn-by-turn directions have become life-changing for those needing directional and situational assistance. Most turn-by-turn directions are conveyed using the device's screen and are often accompanied by audio, but haptic solutions may offer a better alternative to convey this information.

3.4 Virtual

Virtual worlds are a rich part of the modern experience. Whether it be games, simulations, or educational environments, virtual worlds are becoming commonplace with the advent of consumer VR and widespread gaming hardware. One of the issues is that most virtual environments are developed with vision being the primary interaction modality, effectively excluding many individuals from participation. Accessiblity in games is becoming more and more popular. It is no longer uncommon to find color-blind friendly settings in games as well as subtitles and other accessibility features. An example of this in the context of virtual reality is SeeingVR, a suite of VR tools for making VR environments more accessible to people with low vision Zhao *et al.* (2019). Truly non-visual video games though have yet to become mainstream. While some non-visual video games exist, they are few and far between and almost always rely solely on audio feedback. Some of the first non-visual video games were developed for academic purposes such as the Audio-based Environment Simulator

(AbES) games. AbES is a software suite designed to improve real world navigation skills for people with blindness Connors *et al.* (2013). AudioDOOM and AudioZelda SÁNCHEZ and LUMBRERAS (2009) Mirsky (2009) were developed using AbES. AudioDOOM is one such AbES game that discritized a 3D environment into voxels that a user's avatar (and other entities) can move through via adjacent voxels. Users could interact with entities such as monsters by fighting them when in the same voxel, although no aiming mechanics were involved. After playing the game, children were asked to recreate the virtual environment using legos rendering promising results for the development of spatial awareness in the virtual world. In AudioZelda, users navigate a college campus collecting items to develop familiarity with the campus' layout. A more recent serious game for developing spatial skills is called Hungry Cat Chai *et al.* (2019). Researchers designed audio cues users could use for interacting with 3-dimensional maps. The learned layouts were confirmed using physical representations similar to the validation of learned maps in AudioDOOM. A few examples of modern video games accessible without vision are FEER, an "Endless Runner" game Régo (2018) Meyer and Mikesch (2018), Timecrest: The Door, a story-based game with multiple endings and dynamic storylines DMNagel (2017a) Apple (2015) and A Blind Legend, a first person fighting game for both PC and Android Dowino (2019). Timecrest: The Door, is a story game where one's character has the power to control time and their decisions alter the course of the story DMNagel (2017b) Inc (2015). A Blind Legend is an action-adventure game where you fight with a sword and, similar to Papa Sangre, uses a 3-dimensional sound engine to create realistic and immersive soundscapesDowino (2019). One of the most popular audio-only video games was called Papa Sangre and its successor Papa Sangre 2 Barry (2011). While a handful of games can be played with audio only, the majority of video games and virtual environments remain inaccessible to individuals who are blind. All of these environ-

ments were designed to be used without a visual representations from the ground up. Inversely, there have been a few efforts to make visual environments accessible via assistive technology. Rich haptic feedback devices may provide solutions to this problem.

Developers in the Haptics Laboratory of McGill University in 2006 developed a game of "Memory" using the $STRESS^2$ tactile display Wang *et al.* (2006), a more ergonomic version of the original $STRESS$ 1D haptic display Lévesque *et al.* (2005). Instead of images or text to memorize, the "cards" consisted of unique haptic patterns, making for an interesting spin on the classic game of Memory. Likewise, researchers at Arizona State University designed a 2D spatial game based around the Low Resolution Haptic Display (LRHD), a chair affixed with a 4x4 array of vibrotactile motors. The point of the game was to find the goal 2D top-down environment. The user's position was displayed on the haptic chair as well as the goal using unique vibration patterns and the user could move in the environment using a computer mouse peripheral to find the goal. A study using the game found that users were able to learn how to play the game quickly and their performance increased markedly as they played Fakhri *et al.* (2019). An image of the Low Resolution Haptic Display is shown in fig. 3.6. These games are in contrast to audio-only games as they are haptic-only games.

Several devices and systems have been developed as SSDs for virtual environments. Some of these SSDs substitute vision for touch, while others substitute virtual touch for physical touch. For example, in 1998 researchers employed a force feedback joystick called the Impulse Engine 3000 as an interface to virtual textures and objects. Researchers demonstrated a statistically significant relationship between the virtual texture's perceived roughness with the physical analogue and found that participants

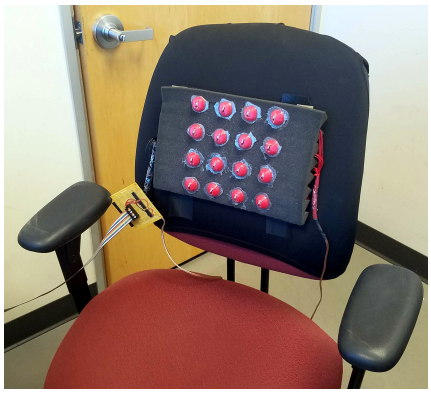


Figure 3.6: The Low Resolution Haptic Display, a 4x4 array of vibration motors mounted vertically on acoustic foam for compliance and damping Fakhri *et al.* (2019)

who were blind were more discriminating than sighted ones using their system Colwell *et al.* (1998). More complex interaction such as discriminating the angle and identity of objects proved more difficult to discern with the system. Researchers found similar results in 1999 using the PHANToM force feedback device (pictured in fig. 3.1) Jansson *et al.* (1999). Again, simple textures were rendered convincingly but the technology was not convincing for object recognition. The primary limitation with these implementations is that only a single point of contact with the "virtual world" is possible, making the interactions akin to poking around with your finger in virtual space.

In response to these problems, researchers proposed non-realistic haptic rendering (NRHR). They argued realistic rendering can be too complicated to parse haptically and non-realistic haptic rendering can make things simpler, giving researchers the chance to eliminate distracting details while emphasizing the important information König *et al.* (2000). To do this, they mapped 3D models onto 2D planes which they argued were easier to navigate. The researchers also propose a different method for guided navigation in virtual environments: a haptic guide. Guiding forces are given to the user as force vectors placed on the PHANToM's stylus König *et al.* (2001).

Similarly, in 2012 researchers using the VTPlayer Mouse developed and tested directional cues via the Braille-like cells . Participants found the cues intuitive and easy to learn Pietrzak *et al.* (2006). This body of research implies that directional guides are useful in navigating virtual environments haptically.

Towards navigating virtual environments "naturally", in 2013 researchers developed the Virtual EyeCane. The virtual cane gives users an auditory signal with respect to the closest object the cane is pointed at in the virtual world Maidenbaum *et al.* (2013), making this system a Virtual Electronic Travel Aid (VETA), similar to the first Laser Cane but unhindered by the limitations of rangefinding in the physical world. A more comprehensive approach was taken by Zhao et al. in 2018 in development of the "Canetroller", which is a virtual cane that gives realistic auditory and haptic feedback in the virtual world so that people who are blind can translate their cane skills to VR. The Canetroller realistically simulates cane forces, impact vibrations, and impact sounds Zhao *et al.* (2018). Besides the EyeCane and Canetroller, there have not been any significant attempts to make accessible to people with visual impairments virtual worlds on equal footing, in essence to take a visual world and present it using an SSD such that they can interact in much the same way as their sighted counterparts. Virtual worlds by their very nature provide mechanisms for making them accessible as object detection and semantic segmentation are less complicated in those environments.

3.5 General Tools for Interacting with Visual Environments

Examples of more modern SSDs include the Social Interaction Assistant, Panchanathan *et al.* (2016), and the VibroGlove, Krishna *et al.* (2010), where facial expressions are identified by the system and relayed to the user via haptics. SSDs

that make use of the auditory modality have also been developed such as KASPA (Kay's Advanced Spatial Perception Aid) Kay (1974), the Sonic Pathfinder, Heyes (1983), and the EyeCane for virtual environments, Maidenbaum *et al.* (2014), and real environments, Chebat *et al.* (2015). More generally, SSDs towards general vision substitution such as the "vOICE", Meijer (1992) and Ward and Meijer (2010), and EyeMusic, Abboud *et al.* (2014), abstract images into tones or musical notes and instruments to convey visual information. Unfortunately, the usability of auditory SSDs for vision substitution is limited as they obstruct a valuable sensory modality (hearing) which is often counterproductive to SS, Krishna *et al.* (2010). Alternatively, haptic SSDs allow the interface to work without obstructing modalities that are often also in use while taking part in typical daily tasks.

One of the most exciting developments in this field is the emergence of Computer Vision methods that are useful for interacting with visual environments. The social media giant Facebook already performs automatic image captioning on uploaded images, updating their alt-text dynamically Metz (2016). The explicitly "assistive" apps Google Lookout and Microsoft Seeing AI give users audio descriptions of scenes captured on a user's phone that are intended to aid in understanding their surroundings, Clary (2018) and Microsoft (2018). Google Lookout describes objects in the scene by giving audio descriptions such as "Trash can, 12-o'clock", but allows the user very little freedom to explore a visual scene in an interactive way. Microsoft's Seeing AI is slightly more sophisticated, augmented with the ability to read text, documents, people, scenes, money, and give illumincation descriptions (color, brightness), Microsoft (2018). While these methods are incredibly encouraging due to the richness of information they provide, their not yet real-time interfaces do not promote intuitive interaction with the visual world. They provide descriptions and summarizations of

visual content, which while impressive and useful in some contexts, hinder a user's agency to explore the visual world deliberately.

One such device that encourages active exploration is the Auditory Night Sight, Twardon *et al.* (2013). Researchers developed a system whereby eye-tracking technology was employed to control what portion of a depth map was relayed via audio to a user's ears (tone depicted depth values). The concept of directing attention via the eyes is compelling: sighted individuals do this intuitively with gaze. But solely providing point-depth cues does little for scene understanding and peripheral awareness. To be truly useful for interacting with rich visual environments, a device must provide real-time feedback, be intuitive and exploratory in nature, and grant the user agency and focus without sacrificing the expansive situational awareness made possible by natural peripheral vision. Combining these very powerful image understanding techniques with a proprioceptive and tactile interace would likely lead to a more effective and meaningful visual environment exploration tool.

Chapter 4

DEEP REINFORCEMENT LEARNING FOR 3D NAVIGATION

The tradeoff between solving the problem for a user and providing them with tools to solve such a problem is one of the first design decisions when developing any assistive technology. For example, a GPS system provides a user a map with a real-time location, while turn-by-turn navigation systems provide a much richer form of assistance. This richer form of assistance often increases the size of the population capable of using a technology, as more of the cognitive load is transfered to the device. However, the same increase in assistance often necessitates an increase in system complexity as well as a decrease in agency for the user. To investigate this tradeoff and the potential of AI-enabled assistive technology to aid in the daily tasks of people with vision impairments, a method for navigational assistance was explored in this chapter.

From navigating the rooms and hallways of one's own residence to navigating a large city, the cognitive functions involved in negotiating an environment to arrive at a predetermined destination are delicate, complex, and in many ways innate. Specialized components of the brain (head direction cells, place cells, grid cells, and border cells) have been shown to be integral to navigation, Moser *et al.* (2015). Although the ability to navigate endows people with independence and self determination, many circumstances can lead to complications in navigation, and a surprising number of people experience such complications. Visual impairments and Alzheimer's Disease are just some examples of common conditions known to cause navigation issues, Duthey (2013) and Monacelli *et al.* (2003).

This is of no surprise, as navigation is a complicated processes requiring multisensory integration over time and space and a strong dependence on memory. Efficiently storing and recalling landmarks and their relationships in space is essential to spacial cognition, and thus navigation, Monacelli *et al.* (2003), leading to large disparities in navigational aptitude, Wolbers and Hegarty (2010). With so many factors affecting navigational ability, there exists real demand for assistive technology in the space of navigational aids. While the advent of ubiquitous GPS has already benefited many with navigational impairments, small scale and indoor navigation remains a challenge. There does though exist promise in the application of emerging computer vision based technologies for navigational aids.

Deep Learning (DL) and Convolutional Neural Networks (CNNs) have recently emerged to solve complex vision-based tasks, Krizhevsky *et al.* (2012); Karpathy and Li (2015). In a reinforcement learning setting, these methods have been shown to learn increasingly complex behavior solely from images, Mnih *et al.* (2013); Bengio (2009); Schulman *et al.* (2015); Mnih *et al.* (2016), from playing Atari games to continuous control. This begs the question: can deep reinforcement learning techniques be employed in assistive technology to aid in navigation? Section 4.1 of this dissertation surveys Deep Reinforcement Learning methods suited for the high complexity of visual navigation, and here a new technique designed for such tasks, GraphMem, is presented. GraphMem's performance is compared so some of these methods in a first-person, vision based navigation task built on the ViZDoom 3D research platform, Kempka *et al.* (2017), shown in Figure 4.1. The findings provide insight into

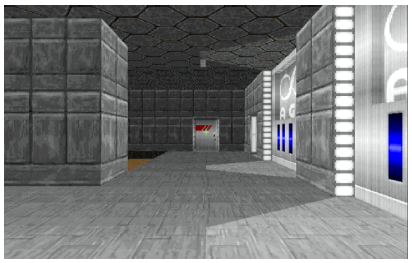


Figure 4.1: Agent's Point of View in ViZDoom.

the difficulties associated with integrating emerging Artificial Intelligence methods with assistive technology.

4.1 Related Work

While Deep Q-Networks, Policy Gradients, and other deep reinforcement learning (DRL) methods, Mnih *et al.* (2015); Bengio (2009), have achieved super-human performance in many domains, some tasks have remained difficult to solve. Especifically difficult are problems with long-term temporal dependencies, Santoro *et al.* (2016), such as navigation. Efficiently solving a first-person maze, for example, requires the ability to memorize where one has been before in order to effectively trim the search space. Failure to do so can result in unnecessary repetition in solving the maze.

Several recent papers have validated the ability of deep networks to make sense of 3D environments using visual information, specifically with a focus on navigation tasks. Supervised methods have been developed such as, Gupta *et al.* (2017), where authors trained a network to infer space through which a robot may travel unobstructed, in order to generate a trajectory for navigating the environment. While there has been success with supervized methods, reinforcement learning paradigms are of predominant interest to our goal, because agent-environtment interaction is integral to navigation. Such approaches can be found in the work of Xie *et al.* (2017),

where the authors used a double-Q network (D3QN) to achieve obstacle avoidance and path planning in a reinforcement learning setting. DeepMind also showed in "Learning to Navigate in Cities Without a Map", Mirowski *et al.* (2018), how natural images can be tamed with CNNs paired with LSTMs in vision-based navigation problems. Researchers in Zhang *et al.* (2017) also explored transfer between navigation tasks, training the model to navigate one environment and subsequently transfering its learning to a new environment in which the walls and objective have been modified. This work is similar to RL^2, Duan *et al.* (2016), a model which achieves a sizable performance increase of 25.5% between the first and second attempts at the same maze. Our task is similar, but with the added complexity of random start positions between the first and second attempt at a maze. While RL^2 was able to store information in its hidden state, it did not make use of addressable external memory. Due to the complexity of spacial navigation tasks in terms of relational connectivity, we chose to explore methods with the capacity for more complex computation: Memory Augmented Neural Networks (MANNs). DeepMind's work in "Learning to Navigate in Complex Environments", Mirowski *et al.* (????), used a stacked LSTM model to solve randomized mazes. While the authors do not employ MANNs in their tests, they stress their applicability to problems of this complexity.

MANNs, sometimes termed "Neural Computers", are characterized by models utilizing an external and addressable memory space, Graves *et al.* (2014, 2016). This allows them to store and recall information relevant to solving problems that require integrating and processing information over time and space more effectively than standard recurrent networks. For this reason MANNs trained in a Reinforcement Learning setting will be the focus of this work. Specifically, we selected the Differential Neural Computer (DNC), Graves *et al.* (2016), and TARDIS, Gulcehre *et al.* (2017), as

MANNs to compare to our model, GraphMem. As a baseline, we also compare to a standard feed-forward multilayered perceptron (MLP) and an LSTM Hochreiter and Urgen Schmidhuber (1997) based model. There has been work on MANNs used in navigation problems: in Oh *et al.* (2016), authors used a MANN, similar to a Neural Turing Machine, Graves *et al.* (2014). The model was tested in a Minecraft-style maze with discrete movement, Johnson *et al.* (2016). Authors also emphasized the use of memory in reinforcement learning tasks in Heess *et al.* (2015), demonstrating the ability of Neural Computers to learn memory-based control tasks. Of these, the most pertinent to the task discussed in this paper is the "water maze" task, in which the agent must first find a hidden objective through random exploration and then subsequently find it again, taking advantage of memories from the initial exploration. Taking inspiration from graph-based representations, Sanchez-Gonzalez *et al.* (2018); Allamanis *et al.* (2018), a MANN is proposed here with graph-like external memory, with the intuition that the spacial connectedness of 3D environments lends itself to a graph-like representation, hence GraphMem.

4.2 GraphMem

GraphMem is a Memory Augmented Neural Network with novel graph-like external memory illustrated in Figure 4.2. The choice of a graph structure for the external memory was inspired by the notion that the strong spacial connectivity of 3D environments would be best represented in memory with strong connectivity. Like most MANNs, GraphMem takes in an observation $\mathbf{x}_t$ at time t from the environment and outputs a distribution on actions $\mathbf{a}_t$ to take at time step $t+1$. The magnitude of the ith element $\mathbf{a}_t[i]$ corresponds to the model's confidence in that action relative to all other actions. Observations are transformed into action probabilities by feeding

the observation into a CNN, producing state representation vector $\phi_t \leftarrow \text{CNN}(\mathbf{x}_t)$. The representation is fed through the Memory Module generating a context vector $\mathbf{c}_t \leftarrow \text{MM}(\phi_t)$. The context vector represents information read from the memory that is relevant to the current observation. The context vector and state representation are then both fed to the policy (a fully connect neural network), which outputs action probabilities $\mathbf{a}_t \leftarrow \pi(\mathbf{c}_t, \phi_t)$. When the state representation ϕ_t passes through the Memory Module, the module reads from and writes to the memory, determining what to store from ϕ_t and where to store it. Information can thus be stored to be recalled when necessary. This process is outlined below.

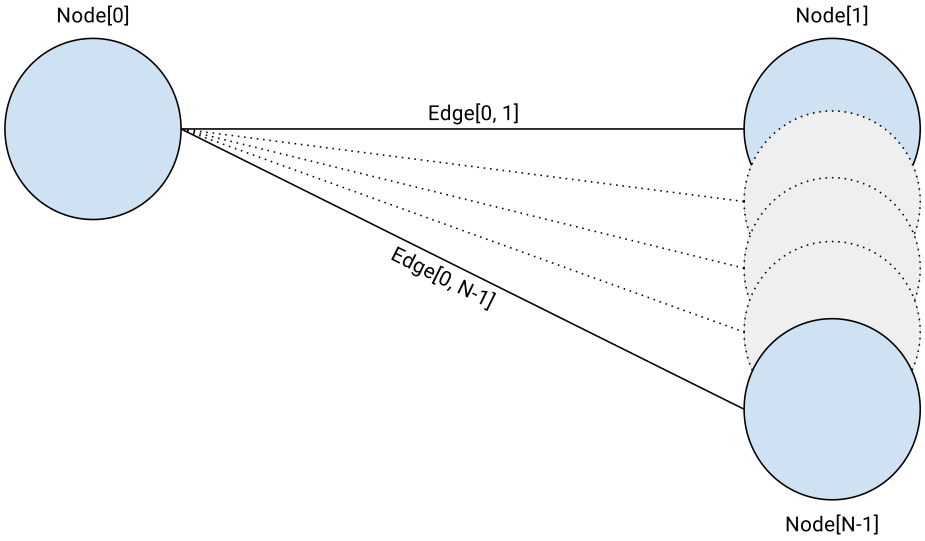


Figure 4.2: GraphMem's memory consists of nodes and edges. Information can both be stored in nodes and edges, encouraging relational reasoning.

GraphMem extends the memory structure of the Neural Turing Machine and its successors,Graves *et al.* (2014, 2016); Gulcehre *et al.* (2017), by modelling external memory as a fully connected graph, illustrated in Figure 4.2, instead of a sequential array. In practice, the memory graph consists of two arrays, one containing the node data and the other containing the edge data. Figure 4.3 illustrates the substructures

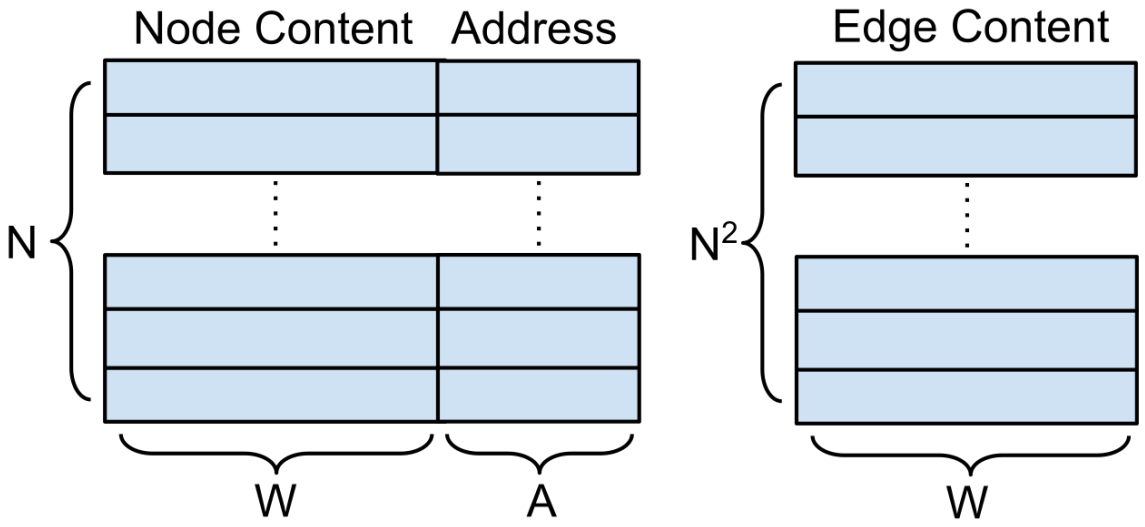


Figure 4.3: Memory architecture: node data array (left), edge data array(right).

of the memory graph. The node array $\mathbf{N}_{\text{ode}} \in \mathbb{R}^{N\times(A+W)}$ is of size $N \times (A + W)$ where N is the number of nodes, A is the address field size, and W is the word size. The edge array $\mathbf{E}_{\text{dge}} \in \mathbb{R}^{N^2\times W}$ is of size $N^2 \times W$, each edge connects a distinct pairs of nodes. The node array's address field is initialized with unique, sparse random vectors. The content field is initialized with zeros, as well as the edge array. At each time step, GraphMem writes to a single graph node and a single graph edge. The node it writes to is based on a content addressing scheme based on content, while the edge it writes to must be the edge connecting the node written to during the previous time step and the node being written to at the current time step. For example, if GraphMem writes to $\mathbf{N}_{\text{ode}}[i]$ at time t and $\mathbf{N}_{\text{ode}}[j]$ at time $t+1$, the edge it writes to at time $t+1$ is $\mathbf{E}_{\text{dge}}[i, j]$. The discrete and graph-like addressing forces GraphMem to discritize its observations and encourages the network to store information relating observations made in close proximity in both time and space in the edges, an ability indispensable to modeling 3D environments.

Similar to TARDIS, Gulcehre *et al.* (2017), we use the Gumbel Softmax reparameterization trick, Jang *et al.* (2017), for discrete memory addressing to retain the ability to differentiate end-to-end. During memory reads and writes, the state representation vector ϕ_t passes through the Memory Module read/write heads (LSTMs) resulting in address logits vector $\mathbf{w}_t \leftarrow \mathrm{RW}(\phi_t)$. This vector describes the categorical probabilities of reading from or writing to a specific node. Equations 4.1 and 4.2 describe how the address logits vector $\mathbf{w}_t$ is transformed into a one-hot vector $\mathbf{m}_t \in \mathbb{R}^N$ describing the memory address of the node to read or write from.

$$\mathbf{g}_t \leftarrow gumbel(\mathbf{w}_t) \tag{4.1}$$

$$\mathbf{m}_t = (one\_hot(argmax(\mathbf{g}_t)) - \mathbf{g}_t) + \mathbf{g}_t \tag{4.2}$$

Equation 4.2 features an $argmax$ and one_hot operation, which are not differentiable. To circumvent this, the gradient only flows through $\mathbf{g}_t$, the last term, bypassing $(one\_hot(argmax(\mathbf{g}_t)) - \mathbf{g}_t)$. This estimates the derivative while allowing backpropagation through a discrete addressing mechanism. Details of the Gumbel Softmax function are described in Jang *et al.* (2017).

4.3 Maze Task

The ViZDoom maze task was designed to reveal how effectively an agent can re-navigate to a location it has been to before, having started at a new location. Figure 4.4 shows a bird's eye view of the map and screenshots of the agent's point of view. Notice, the maze is not "simply connected" as it features detached walls that can fool more simple maze solving algorithms. The goal of this task is to find the "health pack" hidden in the maze. Each episode consists of two phases. For each phase the agent spawns in a random room and must search the maze for a "health pack". The agent

is rewarded, on a per-episode basis, proportionally to the number of steps it takes to reach the goal ("health pack"). The fewer total steps taken (phase1 + phase2), the higher the agent's reward. In both phases of an episode the agent is given the same maze, so that the agent can make use of what was learned about the maze in Phase 1 when looking for the "health pack" in Phase 2. It is important to note that the agent is rewarded in proportion to the summation of steps taken in each phase. The agent will thus learn to minimize the total number of steps and in no way is directed to use its memory to optimize the second encounter. At the conclusion of an episode, both the locations of the "health pack" and furniture in the rooms is randomized, so that the agent must learn a policy that memorizes the maze's composition using its external memory only. This is to prevent the agent from memorizing the maze using the parameters of the model, which is slow and poorly replicates a real navigation scenario.

Figure 4.4: Floorplan of the maze (left) Screenshots of the rooms (middle, right).

The maze environment consists of 9 rooms connected by hallways (shown in Figure 4.4). All of the walls are identical. The only unique features in the rooms are pieces of "furniture" placed in the rooms, one piece of furniture per room. The goal is also placed in a random room at a random offset from the center of the room. This makes seeing the goal from across the maze non-trivial. Because the hallways are narrower than the rooms, furniture and the goal are not necessarily visible from another room.

The agent may also get "caught" on the walls of the room, so the agent must learn efficient movement as well as an efficient exploration policy to maximize its reward. The maze was designed to be non-simply connected, meaning agents that cannot identify and address loop closures may loop indefinitely.

4.4 Training

We trained all of the models using the Asynchronous Advantage Actor-Critic (A3C) algorithm, Mnih *et al.* (2016), which allows for training a model using many distinct instances of the environment in parallel. Parameter updates from the distinct instances are applied asynchronously to a master copy of the policy, which is periodically copied down to the worker copies of the policy that are interacting with the environment. The gradient is describe in Equation 4.3, with policy π, return R_t, value function V, and model parameters θ. The model entropy $H(\pi(\mathbf{x}_t; \theta'))$ is also considered in the gradient to discourage premature convergence to suboptimal policies (scaled by hyperparameter $\beta = 10^{-4}$).

$$\nabla_{\theta'} \log \pi(\mathbf{a}_t|\mathbf{x}_t; \theta')(R_t - V(\mathbf{x}_t; \theta_v)) + \beta \nabla_{\theta'} H(\pi(\mathbf{x}_t; \theta')) \tag{4.3}$$

The models were trained on a 12-core Xeon machine with an Nvidia GTX 1080ti using TensorFlow 1.3.0, Abadi *et al.* (2016). Each model was trained for 30 million time steps ($\sim$ 12 hours). Figure 4.5 shows the training graphs for all models. It is interesting to note that all models show meager performance until 10-15M time steps of training. For our tests, the DeepMind implementation of the Differential Neural Computer was used Graves *et al.* (2016). The LSTM and MLP models used were public A3C, Mnih *et al.* (2016) implementations proven to work on OpenAI Gym benchmark suite environments, Brockman *et al.* (2016). A custom implementation of TARDIS was used as a public version was not available at the time of writing.

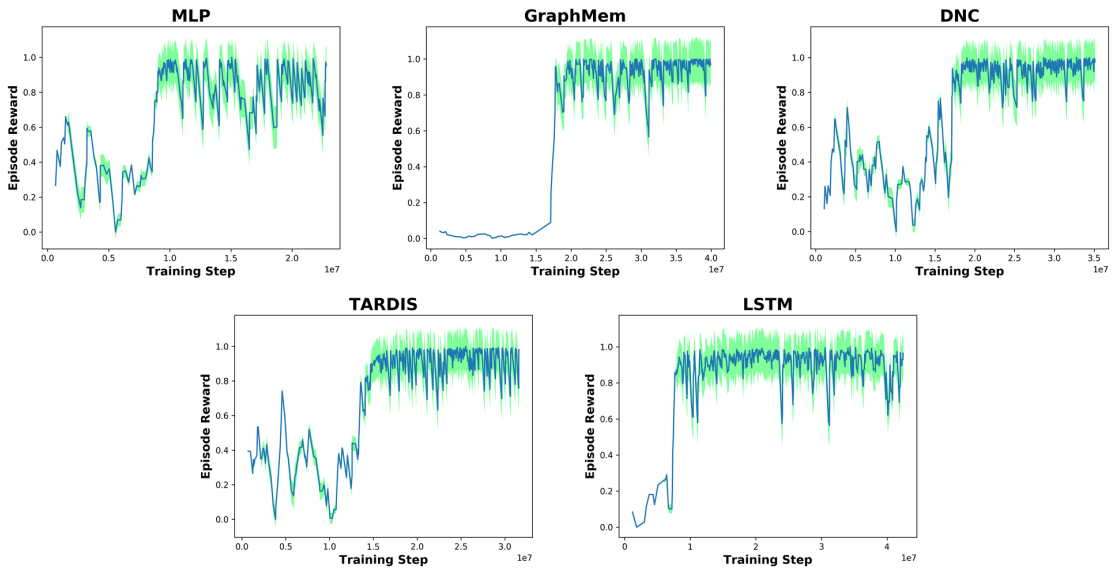


Figure 4.5: Training graphs for all models.

4.5 Results

After training, all models were subjected to 123456 episodes of testing. During testing, the parameters of the network were frozen by disabling backpropagation and the models were subjected to the maze environment for evaluation. Figure 4.6 illustrates the average number of steps taken by the models in solving the maze tasks as well as the percentage improvement of steps between Phase 1 and Phase 2 of an episode. TARDIS and DNC proved to be the fastest models, while GraphMem was the slowest and the MLP and LSTM remain in middle of the pack. With regards to leveraging memory, GraphMem saw the greatest percentage improvement (percentage difference in number of steps between Phase 1 and Phase 2) of all the models, followed by LSTM. It is surprising to note that the two other MANNs were unable to capitalize on having already seen the maze, both models performed about as well as the memoryless MLP model.

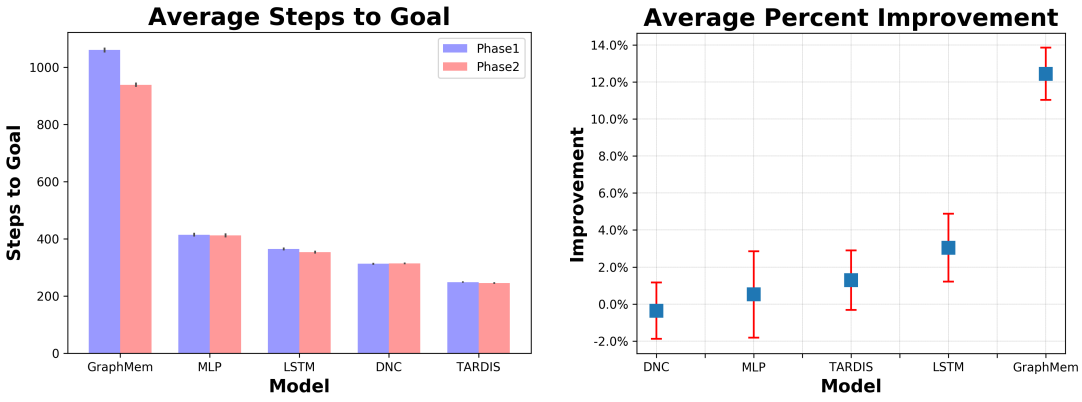


Figure 4.6: Average steps to goal (left) percent improvement from Phase 1 to Phase 2 (right) both with 95% confidence intervals.

	Phase1 $\bar{\mu}_{\textbf{steps}}$	**Phase2** $\bar{\mu}_{\textbf{steps}}$	**Steps Sum**	**Improvement**
TARDIS	248.92 ± 2.82	245.67 ± 2.88	494.59 ± 4.03	$1.29 \pm 1.61\%$
DNC	312.97 ± 3.39	314.06 ± 3.34	627.03 ± 4.75	$-0.36 \pm 1.52\%$
MLP	414.54 ± 6.57	412.27 ± 6.68	826.81 ± 9.37	$0.52 \pm 2.33\%$
LSTM	364.74 ± 4.85	353.56 ± 4.74	718.31 ± 6.78	$3.05 \pm 1.83\%$
Ours	1060.24 ± 8.43	938.76 ± 7.95	1999.00 ± 11.59	$11.45 \pm 1.41\%$

Table 4.1: Results with 95% Confidence Intervals.

4.6 Conclusion and Future Work

Deep Reinforcement Learning methods were applied to the task of vision-based navigation in order to assess the viability of high level solutions for indoor navigation scenarios. While these methods are still in their infancy, this work highlights some encouraging approaches in Artificial Intelligence towards that goal. Upon reviewing first-person video of the trained agents (links: LSTM, TARDIS), some behavioral peculiarities are hard to miss. The videos show considerable redundancy in trajectories and in the case of TARDIS, and odd aversion to turning right preferring inefficient

270° right turns instead. These peculiarities serve as a reminder of the immaturity of these methods and are a sobering warning against implementing them in assistive technology for navigation just yet. Additionally, the complexities of real-world indoor navigation such as changing goals and understanding user intent pose large obstacles to practical implementation. Consequently, technology that conveys pertinant information from the visual world to a non-sighted user via Sensory Substitution is likely a more effective approach.

Chapter 5

LOW RESOLUTION HAPTIC INTERFACE

Conveying visual information to a user who is non-sighted necessitates a Sensory Substitution Device to perform the visual-to-haptic translation. As expounded upon in section 2.4, the skin provides an ample and flexible communication channel for SSDs, and a haptic array can be used to convey a variety of types of information. Most SSDs though are purpose-built, Panchanathan *et al.* (2016), Krishna *et al.* (2010), Eagleman (2014), Novich (2015), and are unfit for scenarios outside of their intended use-case. To rectify this, the Low Resolution Haptic Interface (LRHI) is introduced, a general-purpose haptic interface for sensory substitution that abstracts 2D haptic patterns into "Haptic Images".

While similar haptic displays have been explored in the past, Jones and Ray (2008), such displays have not been validated for interactive applications in user studies. In this section a 4x4 instantiation of the LRHI is evaluated in its effectiveness in conveying abstract information to users as well as for use in interactive applications in the form of a 2-dimensional video game played without vision. Additionally, a form factor that improves upon the response of typical eccentric rotating mass vibration motors as well as a Python interface for haptic displays of this nature are introduced and made public. The Python library to interface with the LHRI can be found here: `https://github.com/bfakhri/lrhi`.

5.1 System Design

Building a general-purpose haptic sensory substitution device required a standardised and general interface. Towards this, we propose that haptic patterns be

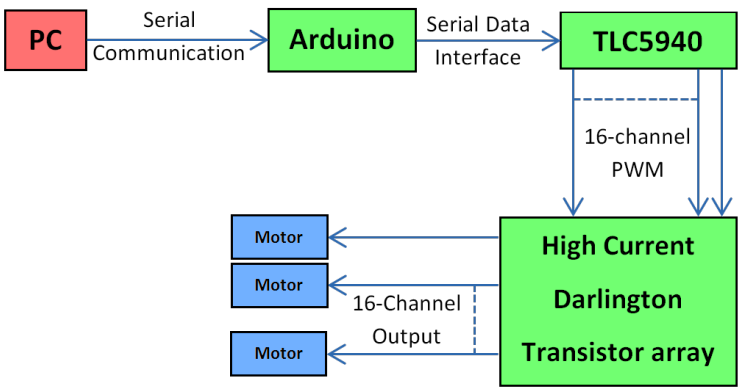
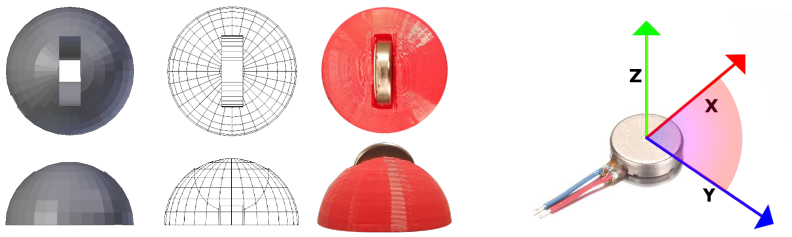


Figure 5.1: Block diagram of the LRHI. In red *Computing Platform*. In green *Controller*. In blue *Display*

abstracted into "haptic images", which are essentially two dimensional arrays of haptic intensities and frequencies $i, f = H[x, y]$ analagous to how a visual image can be modeled as a 2D array of color intensities $r, g, b = V[x, y]$ (RGB model) where x, y are discrete coordinates relating to space. A series of haptic images can thus convey moving patterns over time similar to how a series of images becomes a video. The LRHI is a system that communicates using "haptic images" and converts them into tactile representations. The LRHI consists of a *computing platform* which sends haptic images to be displayed, a *controller* which intereprets the haptic images and converts them into anolog signals, and a *display* which converts the analog signals into vibrotactile actuation. Figure 5.1 shows a block diagram of the LRHI.

The *computing platform* can be any USB enabled computer: its role is to generate the haptic images in a digital and abstract form. The computing platform may take on a variety of roles in generating the haptic images. In sensory substition applications for instance, the computing platform converts images from a video stream into haptic images and sends them to the controller. The actual algorithms for conversion are left up to the designers. In our incarnation of the LRHI, it communicates with



(a) (b)

Figure 5.2: Motor Housing: (a) digital and 3d-printed models (b) vibration axis

the *controller* by sending 4x4 8-bit haptic images.

The *controller* consists of an Arduino microcontroller, TLC5940 analog to digital converter, and a collection of high-current Darlington Transistor Arrays. The Arduino accepts the haptic image and using the TLC5940 converts the haptic image into 16 analog electrical signals (8-bit PWM). These are transmitted to the transistor arrays where the signals are amplified and made suitable to drive the *display*. A full version of the LRHI would allow haptic images to specify not only an intensity but also a vibration frequency for each actuator on the display.

Our prototype of the *display* consists of a 4x4 array of pancake motors housed in custom 3D printed mounts that orient the motors orthogonal to the user's back. The housing is shown in figure 5.2a. This accomplishes two objectives: first, the vibration axis is made perpendicular to the user's back (illustrated in figure 5.2b). Second, the contact point is made smaller. These two objectives increase the perceived intensity of the vibrations which is especially important when the user is wearing thick clothing. The motors and housing are mounted on accoustic foam to provide a maleable surface that adheres to a user's back and simultaneously transmits minimal intermo-

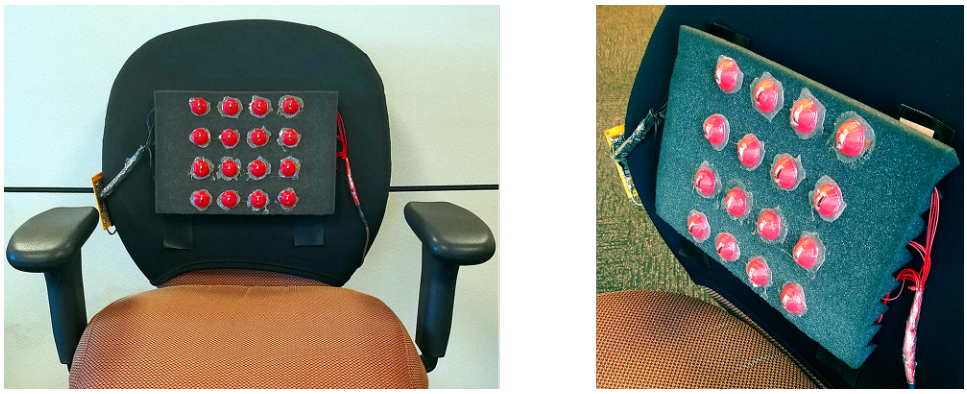


Figure 5.3: 4x4 haptic display mounted on an office chair

	Current (A)	**Power(W)**
Idle	0.05	0.17
Single Motor	0.10	0.33
1/4 Motors	0.19	0.63
2/4 Motors	0.28	0.92
3/4 Motors	0.34	1.12
4/4 Motors	0.41	1.35

Table 5.1: Power consumption characteristics of the Low Resolution Haptic Display

tor vibration. The haptic display is shown in figure 5.3.

The haptic display consumes 50mA in an idle state with a maximum consumption of 412mA when all motors are at full power (energy consumption summarized in table 5.1). During the non-interactive portion of the user study, the LRHI had a mean power consumption of $0.73W$. During the interactive portion of the study the LRHI showed a mean power consumption of $0.56W$.

5.2 User Study

In order to assess the LRHI's potential as an SSD, a preliminary user study with 8 participants was performed to explore its ability to convey information through haptics. The study consisted of a non-interactive and an interactive component. The non-interactive portion consisted of 3 phases wherein participants were introduced to a finite set of haptic patterns during "familiarization" (being exposed to each individual pattern only once) and were asked to recall those patterns during "testing".

During the non-interactive testing portion of the study participants were given the option to repeat the pattern if they were not confident in their assessment. Phase 1 consisted of static patterns (Top Left, Bottom Right, etc). Phase 2 consisted of patterns that vary across space and time (Left to Right, Top to Bottom, etc). Phase 3 is similar to Phase 2, but users were asked to recall how fast the pattern was displayed (Left to Right - Fast, Top to Bottom - Slow, etc) in addition to the original pattern identity (Left to Right). The patterns increased in complexity in each subsequent phase, beginning with simple single-motor patterns to patterns that move through space and time. Participants were given the option to repeat a pattern if they were not confident in their initial assessment. The patterns for each phase are illustrated and described in section B.

In order to assess the LRHI's potential in interactive environments, we designed a completely haptic, cat-mouse game to play (illustrated in figure B.3). The user plays as a cat, and the goal is to find a mouse. The cat is presented on the haptic display as a solid vibration, while the mouse is a pulsing vibration. Participants used a computer-mouse to control the position of the cat on the haptic display, leading it

Non-interactive	**Repeat %**	**Error %**
Phase 1	1.13%	0.92%
Phase 2	1.46%	0.83%
Phase 3	2.60%	3.12%
Total	1.73%	1.62%

Table 5.2: Results for Non-interactive Phase

towards the mouse - the goal being to catch the mouse as quickly as possible. The duration between the beginning of the game and capturing the mouse was recorded, each participant playing 60 games in total (results shown table 5.3). Increasing performance in this game (decreasing game time) was intended to show that participants were in fact able to learn to use the LRHI to interact with dynamic environments.

5.3 Results

For the non-interactive portion of the study, participants were able to identify the patterns with considerable accuracy. Phase 1, which included static patterns only did not significantly differ in accuracy over Phase 2 (dynamic patterns). Only when participants were asked to discern both the pattern and the speed at which it was presented did performance suffer slightly. Results are compiled in table 5.2 - participants were able to achieve an aggregate accuracy of 98.38%.

For the interactive portion of the study (illustrated in figure B.3), participants were able to capture the mouse in 4.81 seconds on average, and showed a significant performance increases the longer they played. A comparison of the first third of the

Interactive	Avg	StdDev
Total	4.81 s	2.99 s
First 3rd	6.64 s	3.23 s
Last 3rd	4.40 s	2.78 s

Table 5.3: Results for Interactive Phase in game times: Lower is better.

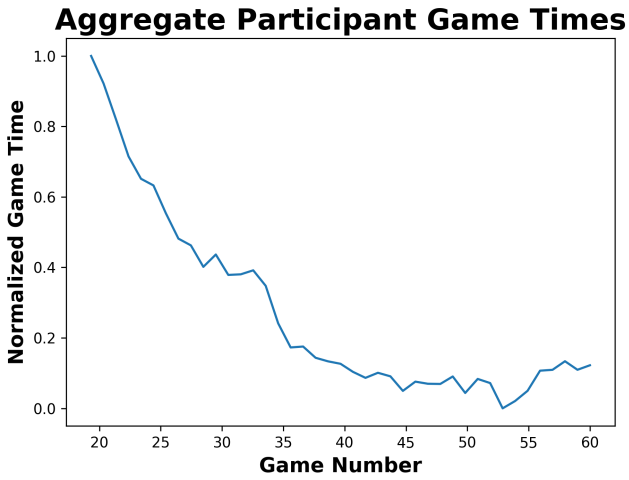
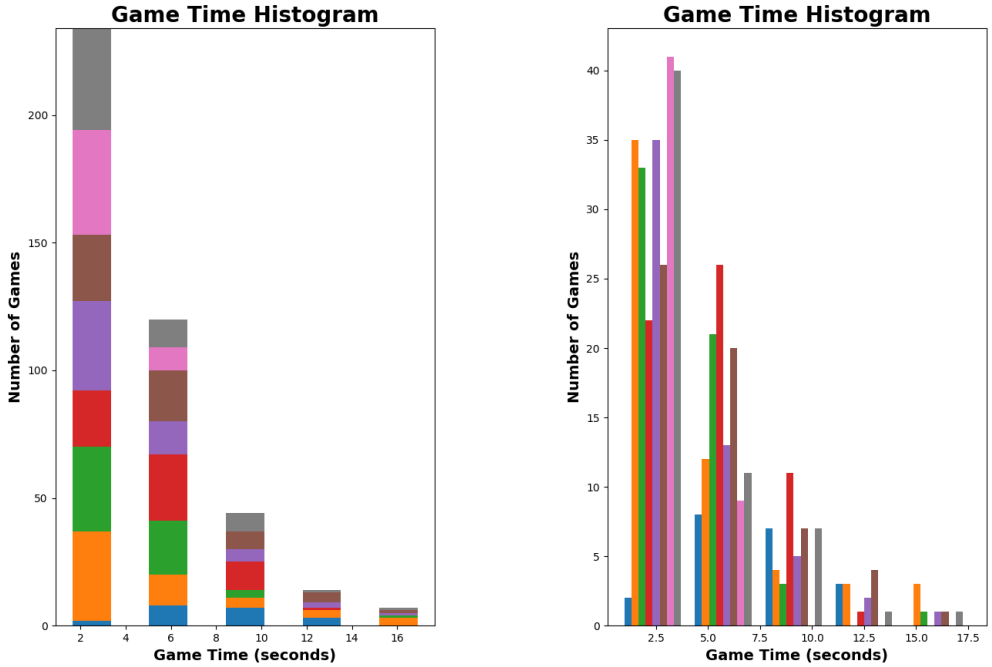


Figure 5.4: Normalized mean game times of all study participants over time.

gaming session (first 20 games) and the last third as well as total performance can be seen in table 5.3. Figure 5.4 illustrates the participants' performance over time.

5.3.1 User Feedback

Overall, feedback was positive regarding the playability of the game with the display. Participants reported the display being strong and vibrations clear and easily felt through clothing, one user even stated that the vibrations were too strong. On the critical end, some participants reported that the mouse was sometimes confused with the cat in the game, especially if they were instantiated near each other. In this scenario, the player would find themselves attempting to move the wrong entity and



(a) (b)

Figure 5.5: Histograms of the game times color coded by participant. (a) Stacked game times (b) Individual game times

resulted in a significantly longer game (lower performance). This is echoed in the data: a histogram of the game times (figure 5.3.1) shows that almost all participants had a game or two in the ~ 15 second range, even the most performant participants. This confusion could possibly be mitigated by coupling a vibration pattern on the hand syncronized with the vibration pattern of the entity they are controlling.

5.4 Conclusion and Future Work

In conclusion, a general purpose haptic interface for Sensory Substitution in interactive settings was introduced and evaluated with a user study. The performance

of the interface with respect to non-interactive scenarios was shown to be in-line with similar haptic displays, Jones and Ray (2008). Additionally, the performance in interactive settings is promising: participants were able to consistently improve their gameplay during the first 35 games and subjective user feedback was encouraging. The results bode well for the application of the LRHI for other spatial interactive scenarios, as participants were able to intuitively grasp the spatial nature of the cat-mouse game. The occasional confusion between the cat and mouse entities in the game described in section 5.3.1 though imply a limit to the complexity of entities shown spatially on the display: the display is well suited to conveying spatial information but not necessarily so for a variety of entity types. For this reason, generalizing to 3 dimensions will likely requires a more intricate solution in order to effectively convey the pertinant visual information in the more-complex 3D visual environment and the real-world.

Chapter 6

FOVEATED HAPTIC GAZE

Virtual worlds are becoming ubiquitous as digital technology permeates society, with augmented and virtual reality being the latest and most immersive manifestations. Unfortunately, the visual domain is central to most virtual worlds, making them inaccessible to people with visual impairments. People with visual impairments already face accessiblity hurdles when using technology but virtual worlds remain one of the most inaccessible mediums. Two competing approaches exist to correct this dilemma. Designers of virtual worlds develop the environments with accessibility in mind in the first approach. Secondly, accessiblity engineers develop tools to make existing virtual environments accessible. While the first approach is gaining traction and public awareness, developers of virtual environments seem to have been excused of this responsibility as accessible virtual environments remain extraordinarily scarce. The second approach has the potential to affect many existing environments. An example of the effectiveness of the second approach is screenreader technology. Screenreaders made digital text and many of the invaluable capabilities of smartphones accessible to millions of people with visual impairments. Additionally, tools that aid in the understanding of 3-dimensional virtual worlds would be best poised to generalize to real-world visual problems faced by people with visual impairments.

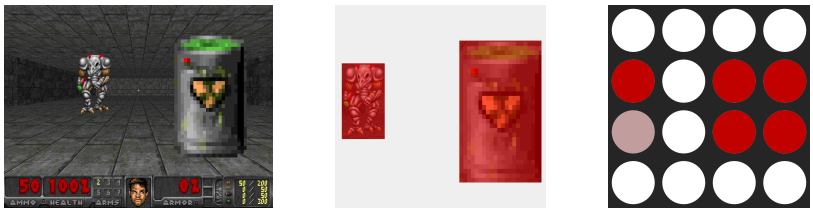
In this vein, this chapter describes the development and assessment of a transformative technology for interacting with 3D visual environments entirely through haptics. "Foveated Haptic Gaze" (FHG) is a concept embracing the characteristics of the human visual system that make it so well-suited for interacting with 3-dimensional environments (elaborated upon in section 2.1. FHG makes use of an attentional

mechanism similar to foveated vision that allows users to focus on objects while simultaneously allowing for peripheral awareness. Foveated Sensory Substitution has also been explored by Capelle *et al.* (1998), who developed a vision-to-audition SSD with a higher resolution portion of the field of view in the center. These researchers did not though separate the peripheral and attentional portions of vision as with FHG. This combination gives users the ability to explore an environment in detail while maintaining broader situational awareness, making "Foveated Haptic Gaze" one of the only vision-to-haptic interfaces flexible enough to generalize to the real world.

To validate this approach, a first-person shooter game based on Doom was developed as well as a working prototype of the Foveated Haptic Gaze system. The system was then evaluated in a user study with both individuals that are sighted and individuals with visual impairments for usability and effectiveness as an SSD for interactive tasks. Seeking to develop an approach that is useful to people with limited or no sighted priors, the user study measured the in-game performance of both populations to understand the effects sighted priors may have on the effectiveness of FHG.

6.1 Method

Human gaze is characterized by aligning the optical axis of the eye to whatever in the visual field one is interested in. The optical axis also happens to be aligned with the fovea, an area of the retina featuring the highest density of photosentitive receptors, Hudspeth, A.J.; Schwartz, James; Siegelbaum, Steven; Kandel, Eric; Jessell (2012). Gazing is thus directing one's visual attention by aligning the most acute portion of the retina with whatever is of interest. The rest of the retina is responsible for peripheral vision, enabling a wide (up to 220° horizontally) spatial awareness in direct spatial relation to one's focus, Szinte and Cavanagh (2012). Thus the human



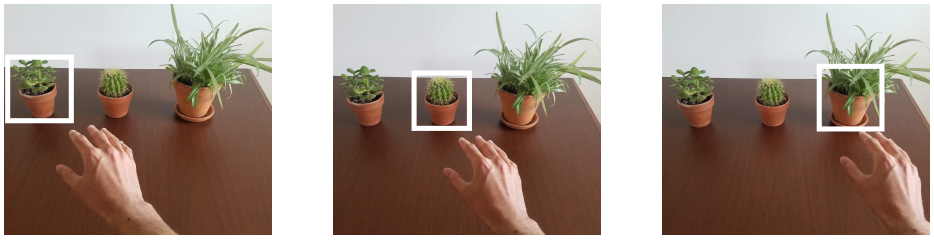
(a) (b) (c)

Figure 6.1: (a) Original image of room (b) Objects of interest highlighted (c) Corresponding motor array activations

visual system has the capacity for high resolution as well as expansive field-of-view thanks in part to foveated vision.

6.1.1 Foveated Haptic Gaze

We borrow the concept of foveated vision to develop a biologically inspired haptic implementation called Foveated Haptic Gaze (FHG). In the same way sighted individuals gaze with their eyes by pointing their foveas at objects of interest, using our system individuals with visual impairments can gaze in a visual environment by pointing their hand at objects of interest (an illustration can be seen in figure 6.1.1). The user wears a purpose built haptic glove (shown in figure 6.1.3) and when they point their hand at an object, details of the object are haptically conveyed via the glove equiped with vibration motors on the finger tips. This provides an analog to the high-resolution fovea, while a back-mounted haptic display (shown in figure 6.7a), Fakhri *et al.* (2019), endows the user with peripheral awareness (Haptic Peripheral Vision) of their entire field-of-view. The system thus partitions a user's experience into two channels: one for high-fidelity and one for wide field of view. The back display alerts the user to the presence and coarse location of objects (obstacles, doors, persons, etc) while pointing a hand towards these objects provides the user with finer



(a) (b) (c)

Figure 6.2: User's hand position determines where they are gazing: (a) Gazing at leftmost plant (b) Gazing at middle plant (c) Gazing at rightmost plant

details of the object, such as the object's identity (e.g. "door", "person", etc). To integrate these two systems so that a user can relate the position of their haptic gaze with their haptic peripheral vision, the system displays the position of their gaze with respect to their field-of-view on the back display. Practically, a user feels on their back where objects are and where their gaze currently is, moving their hand to align these indicators is essentially gazing at the object. This is akin to noticing an object in your periphery then gazing at it for more details. In order to capture a user's "Haptic Gaze" (where the user is pointing their hand), a Leap Motion Controller was used. See appendix section A for design considerations using this device. To illustrate the effectiveness of our approach we created a gaming environment with which participants can interact with rich 3D spatial situations.

6.1.2 Gaming Environment

The First-Person Shooter (FPS) genre of video games was a natural choice for testing the system's efficacy because FPSs offer a realistic simulation of the first-person experience as well as mechanics like aiming and shooting that require keen visuospatial awareness to play effectively. The game DOOM is one of the most iconic and modded FPS games in existence, making it our choice for developing experimental

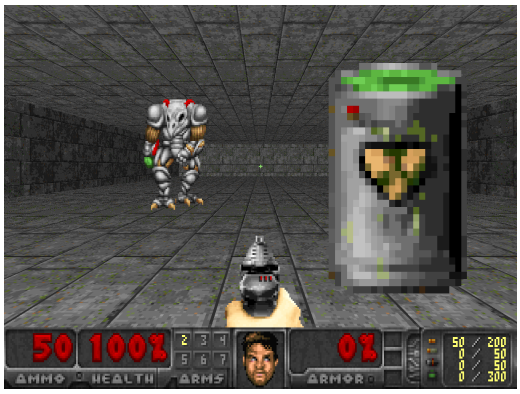


Figure 6.3: Doom Environment featuring a "Hell Knight" monster on the left and explosive barrel on the right.

environments using the ViZDoom platform. ViZDoom, Kempka *et al.* (2017), enabled us to develop visually rich, low-overhead, and responsive DOOM environments for use in our experiments. A system that can empower users to effectively play a game like DOOM has the best chances of generalizing to real-world interactive visual environments. Figure 6.3 shows an image of the environment from the first person perspective.

We designed a level consisting of 10 connected rooms. The player runs through the rooms encountering monsters and explosive barrels (shown in figure 6.3). Figure 6.4 shows a top-down view of the rooms: there are 11 monsters and 5 explosive barrels randomly positioned in the rooms, with more monsters/barrels occuring in later rooms. The objective is to shoot as many monsters as possible while not shooting the explosive barrels. The player's score is the difference between the number of monsters killed and the number of explosive barrels shot: $score = monsters - barrels$. A user willl feel the presence and position of monsters or barrels in their field of view on their back via the haptic display. To ascertain whether the objects are monsters or barrels, the user must gaze over the object with their hand.

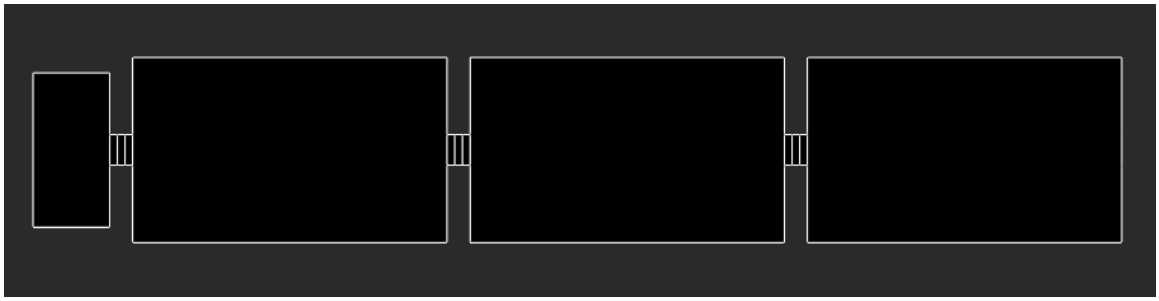


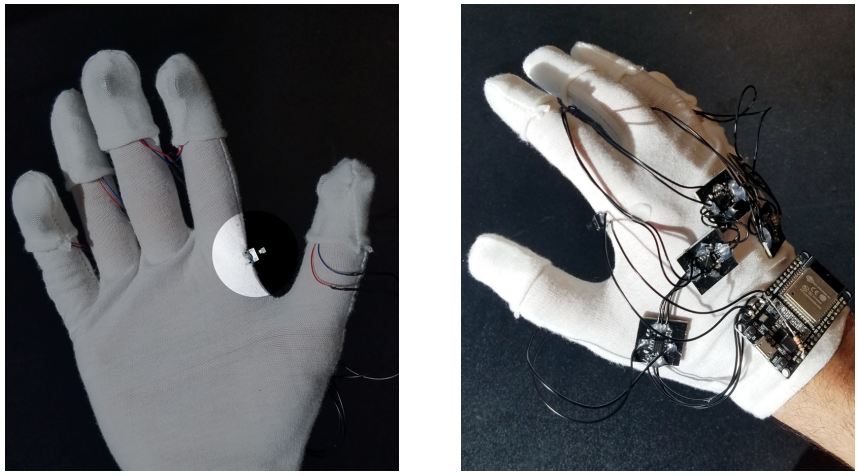
Figure 6.4: Bird's eye view of the (abridged) game map used in the study. The full map consisted of 10 interconnected rooms.

6.1.3 System Design

A user wears a glove equiped with a button and vibration motors on the finger tips (shown in figure 6.1.3). The vibration motors convey information about what the user is gazing at, and in the case of our hallway game, reveal to the user whether they are gazing at a monster or a barrel. The user's hand position is tracked with a Leap Motion Controller, and the 3D coordinates of the hand are mapped onto the field of view of the player's avatar. The location of objects in the avatar's field of view is extracted from the ViZDoom environment and is mapped, along with the user's gaze position, onto the haptic display on the user's back. A diagram of the whole system can be seen in figure 6.6 and a video demonstration of gameplay is available using the link: `https://youtu.be/59-l8B2Xq4E`.

6.1.4 Experimental Design

Five participants with visual impairments and ten sighted participants were recruited for the user study. At the beginning of the study, participants were acquainted with the hardware they would be using: haptic display (chair), haptic glove, and Leap Motion Controller. Participants were then introduced to the concept of FHG by performing an introductory exercise that activated the Leap Motion Controller and



(a) (b)

Figure 6.5: (a) Pancake motors and button (highlighted) on haptic glove (b) Image of the back of the glove with motor driving hardware shown

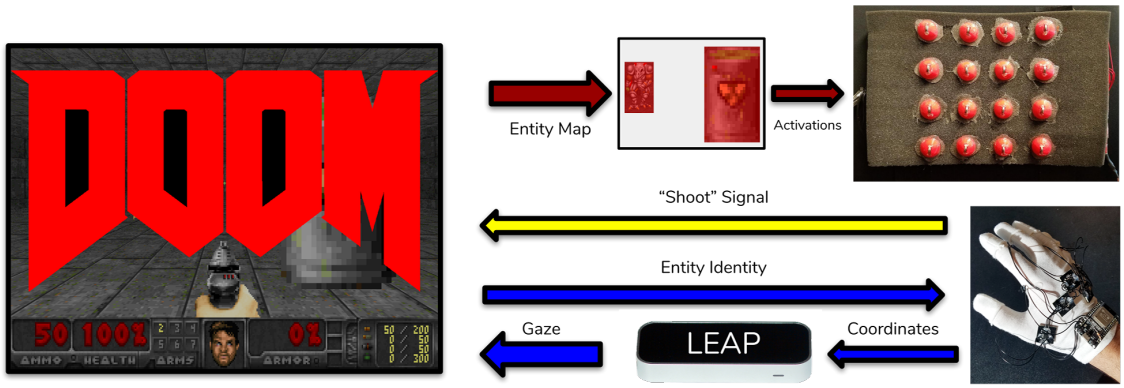


Figure 6.6: Diagram of experimental setup. The ViZDoom game engine sends an entity map to the haptic display (red) to be felt by the user. The hand's movements are tracked by the Leap Motion Controller and its position is converted to gaze coordinates on the avatar's field of view. If the gaze intersects with any entities their identity is sent to the glove (blue). If the user presses the trigger button on the glove a shoot signal is sent to the ViZDoom environment and the avatar fires in the direction the user is gazing.

haptic display only. The participant's hand was tracked and displayed on their back using the haptic display and participants were encouraged to acquaint themselves with the limits of their field of view. The purpose of this exercise was to illustrate the mechanics of the gazing mechanism e.g. moving one's hand to the left moved their gaze to the left on their back. Next they were introduced to the concept of Haptic Peripheral Vision.

Users' avatars were placed in a room in the ViZDoom environment populated by one monster and one explosive barrel on either size of their field of view. The haptic chair relayed the locations of the monster and barrel to them by pulsating on their backs (see figure 6.1c). The location of their gaze was also conveyed by the haptic display via a solid vibration; consequently users learned to gaze towards the objects in the room by aligning the gaze vibrations with the pulsating "entity" vibrations on their back. Upon placing their gaze over one of the entities (monster or barrel), the identity of the entity was conveyed to the user via the glove's vibration motors in a coded manner. Users were instructed to discriminate a "monster pattern" and "barrel" pattern. After exploring the room by gazing over the entities, participants were instructed to shoot both entities. When the barrel is shot it explodes and creates a load explosion sound while shooting the monster results in a triumphant "winning" sound. These are the only audio cues in the whole game other than rhythmic game music.

After this explaination, participants were asked if they were comfortable with the interface and objective and were given the chance to enter the demo room once again, after which the experiment began. Participants entered the hallway game environment described in section 6.1.2 and illustrated in figure 6.4 to score as many points as possible. Participants were asked to play 7 games (each taking about 1.5

(a) (b)

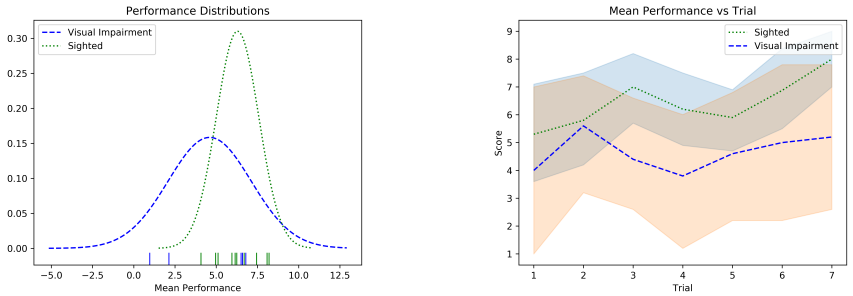
Figure 6.7: (a) Haptic display on office chair (b) Closeup of motor array

minutes to complete) and their performance as well as auxiliary metrics (shots fired, hits, misses) were recorded during their gameplay.

6.2 Results

To assess playability as well as any differences in usability between sighted and users with visual impairments, we measured a player's score throughout every game played. On average, sighted users obtained higher scores although the majority of users with visual impairments also clustered towards the center of the sighted performance distribution shown in figure 6.8a. Both populations saw an initial increase in performance although sighted individuals maintained an upward trajectory slightly longer while participants with visual impairments leveled off sooner. Figure 6.8b illustrates their performance over time. The theoretical maximum score is 11 as there are 11 monsters to destroy, although their positioning often makes them difficult to destroy due to their brief visibility.

To assess a player's ability to make decisions on-the-fly, they were instructed to avoid shooting explosive barrels, as it would negatively impact one's score. These mistakes as well as good hits and complete misses were recorded on a per-game basis. Players overal made few mistakes, many averaging below one mistake per game (figure



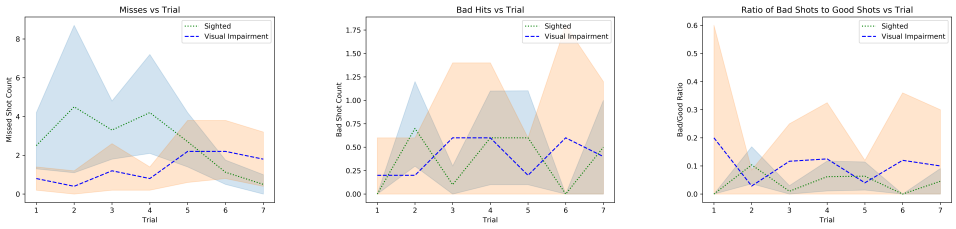
(a) (b)

Figure 6.8: (a) Normal distribution fit to the performance of both participant populations averaged over all trials (b) Performance over time averaged over participants

6.9b), indicating that the glove feedback was clear and intuitive: as a ratio of mistakes to good shots (monsters killed), most players stayed below 1/10.

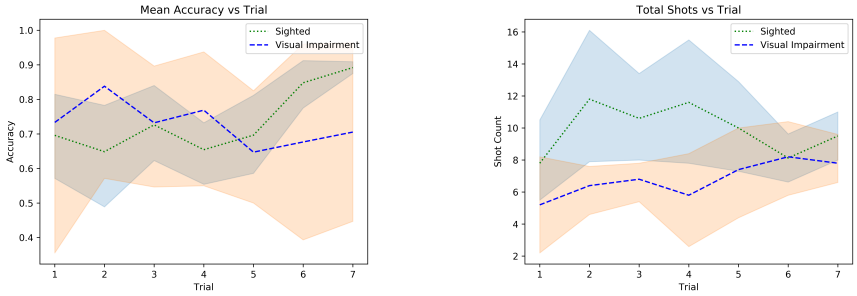
Participants with visual impairments initially missed less than sighted participants, trending upwards throughout the trials eventually ending slightly higher than sighted participants (figure 6.9a) . Inversely, sighted participants missed more often from games 1 through 5, but during the last two games ended with slightly fewer average misses. These trends imply that participants with visual impairments tended to approach the game more cautiously than sighted individuals, becoming more comfortable as games went on while their sighted counterparts were more cavalier to begin with and reigned in their enthusiasm as the games progressed. This is supported by the total shot counts per trial figure, plotted in figure 6.10b, where it can be seen that sighted participants initially took many more shots than those with visual impairments.

Both sets of participants performed similarly with regards to accuracy (hits over total shots taken per game) as illustrated in figure 6.10a. Players achieved an accuracy between $70 - 80\%$ during the first 5 games, indicating that the aiming and gazing



(a) (b) (c)

Figure 6.9: (a) Misses per trial (b) Shots that hit a barrel per trial: mistakes made by participants by shooting an entity they were instructed not to shoot (c) Ratio of enemies killed to explosive barrels (mistakes) over trial. There is a large variance in performance initially for participants with visual impairments that quickly dwindles as the participants learn from their mistakes.



(a) (b)

Figure 6.10: (a) Accuracy over time (hits / shots taken) averaged over all participants in each populaiton (b) Total shots taken per trial, averaged over all participants in each population

mechanics of the system were usable for real-time interactions. Interestingly, sighted players' accuracy rose to touch 90% during the last two games, in tandem with their dip in misses (figure 6.9a).

6.3 Conclusion and Future Work

Results from our user study indicated the playability of the Doom game was maintained without vision as most participants were able to achieve respectable performance metrics and accuracies. This was supported by positive subjective user feedback with regards to the system design. Differences in performance between test groups were small, boding well for our approach having only slight sighted usability bias. A more extensive analysis is required to rule out a sighted performance bias and may inform design decisions to make the approach even more intuitive to people without vision. The results also indicate that individuals with visual impairments approached the game more cautiously, becoming less cautious over time while sighted participants approached the game with less caution and became slightly more cautious over time. Consequently, future approaches may benefit from designs that encourage confidence inspiring exploration. Furthermore, the presentation of peripheral vision information (Haptic Peripheral Vision) can likely be improved. The accuracy assessments indicate that foveated gaze feedback worked well, while destroying all 11 monsters remained difficult for both populations, as brief appearances of monsters were sometimes missed. A higher resolution haptic back display or one with wider back coverage such as the HaptWrap, Duarte *et al.* (2019), may mitigate this by providing more salient peripheral awareness feedback.

To generalize to the real world, it is imperative that the system be capable of detecting objects in images that the user has deemed "of interest". In virtual environments, bounding boxes for entities are often conveniently available, as is the

case in ViZDOOM. Real-world visual scenes almost never offer the same convenience. Coupling Foveated Haptic Gaze with a computer vision method for learning person specific "objects of interest" would allow the system to be generalize to a variety of real-world scenarios.

Chapter 7

ONE-SHOT OBJECT DETECTION FOR PERSON CENTERED VISION

Object Detection methods have come a long way in a short amount of time, seeing exponential-like growth in performance during the last decade Zou *et al.* (2019). Current methods feature both high mAP as well as high framerates on modern hardware: performance has reached a point such that applying these techniques to real-time interactive use-cases is within reach. A particularly straightforward and beneficial use for these methods is aiding those with visual impairments find objects in their surroundings more effectively. In order for this to be practical, the methods must be capable of detecting objects users are actually interested in. Often a user is interested not in finding any item of a certain class, but a specific item of that class. In other words, a user is not likely interested in finding any water bottle, but instead *their* water bottle. If a user with a visual impairment is looking for *their* bottle, using an object detection model trained on the COCO dataset with class "bottle" is not of much use: figure 7 illustrates this conundrum.

Training an object detection model for a new object of interest, such as the bottle featured in figure 7.1a would require the user to generate a dataset large enough to train the model. Few-shot methods reduce the number of training samples for the novel class needed, but they still require hand labeled images of the novel class. This of course is impracticle and such an effort would likely overcome the convenience a personal object detector would provide to a person with a visual impairment. This begs the question: can modern object detection methods achieve competitive performance on user-defined objects of interest with minimal user effort? In this section a data collection and augmentation method called 1SODDA (1-Shot Object Detec-

(a) (b)

Figure 7.1: (a) User specified "object of interest" bounded in white box (b) Output of YOLOv3 trained on the COCO dataset: notice the non-specificity of the attribution of class "bottle"

tion by Data Augmentation) that attains competitive performance with minimal user effort is introduced.

7.1 Literature Review

Currently the most competitive object detection methods are region-based methods. Region-based object detection methods such as RCNN, Girshick *et al.* (2014), employ a "region proposal generator" which generates region proposals in images. In RCNN, those regions are reshaped and fed to a convolutional neural network to extract a feature vector describing the region. This feature vector is used to train a single linear SVM for each class. This method proved highly accurate, achieving a mAP of 53.3% on the VOC 2012 dataset, becoming the state-of-the-art method in 2014. While highly accurate, the method is computationally expensive mostly due to the number of region proposals considered by the network (about 2000 per image): inference on a single image can take as long as 47 seconds on a modern GPU. For use-cases where real-time feedback is necessary such as searching for an object in a

visual environment, this is too slow. These speed issues were partially addressed by Girshick with a version of RCNN called Fast-RCNN, Girshick (2015). In Fast-RCNN, the input image is passed first through a CNN to give a feature map. Region proposals are then made on the feature map instead of the original image and pooled using a region-of-interest (RoI) pooling layer to give a fixed size feature vector for each RoI. These feature vectors are then sent to a fully connected portion of the network to predict the classes and final bounding boxes. Performing the region proposals in the feature space saves significant computation: the author claimed a speed up of 10-100x for inference. Even faster approaches such as You Only Look Once (YOLO), Redmon *et al.* (2016), where the bounding boxes, classes, and confidence scores are predicted from a single pass of the network. YOLO resembles an autoencoder, with an encoder and decoder network, but the decoder predicts bounding boxes, classes, and confidences instead of the pixels of the input. The latest, most accurate, and fastest manifestation of YOLO is YOLOv3, Redmon and Farhadi (2018), which partitions the prediction into 3 scales (small, medium, large) so that the bounding boxes of very small or very large objects are predicted more accurately. YOLOv3 achieves respectable mAP on the COCO dataset, at 55.3% at 35 frames per second. While these object detection methods boast great performance (some versions of these models are capable of running on mobile platforms), they require immense datasets with many training examples of each target class to train. To be useful for finding specific objects of interest with minimal user effort though, a one-shot learning approach must be taken.

Few one-shot learning methods for objects detection have been explored before, Biswas and Milanfar (2014), with some success. More recently, deep methods for one-shot learning have become available such as RepMet, which employs a distance learning framework on top of standard object detection models to render competitive

mAP in one-shot scenarios (42.2%), Karlinsky *et al.* (2019). To newer methods have emerged as promising approaches to few-shot learning for object detection: meta-learning and fine-tuning. Meta-learning methods seek to learn and exploit the structure and relationships within and between classes to perform a task. An example of this was proposed by Fan *et al.* (2019), which uses a model similar to a Siamese Network, Koch *et al.* (2015). A similar method employs a "re-weighting module" which modifies vectors given by a shared feature extractor given a query image, Kang *et al.* (2020). Fine-tuning approaches can be partitioned into two categories: jointly fine-tuning and entire model fine-tuning, Wang *et al.* (2020). In "jointly fine-tuning" approach the model trained on a mixture of samples from the novel and base classes. Similarly, an "entire model fine-tuning" is characterized by first training the model on the base classes and afterwards training on a balanced dataset of base classes and novel classes. The most competitive fine-tuning method to date is a refinement of the "entire model fine-tuning" approach, called "two-stage fine tuning", introduced by Wang *et al.* (2020). This method initially trains a model on the base classes and in the second stage trains solely the last layer on a balanced combination of novel and base class data. Two-Stage Fine-Tuning achieves state-of-the-art performance on most of the few-shot object detection benchmarks, Wang *et al.* (2020).

These one-shot learning methods still require at least one hand annotated image. To be most useful in a person-centered context the ideal method automatically generates the single data sample used in the one-shot learning scenario. None of these techniques have that capacity. In this chapter a unified, minimal effort one-shot learning method for object detection based on a novel data collection technique paired with a synthetic dataset generation process is proposed. This method can be used to learn an object detector for personal objects for which no publically available annotated datasets exist.

7.2 Method

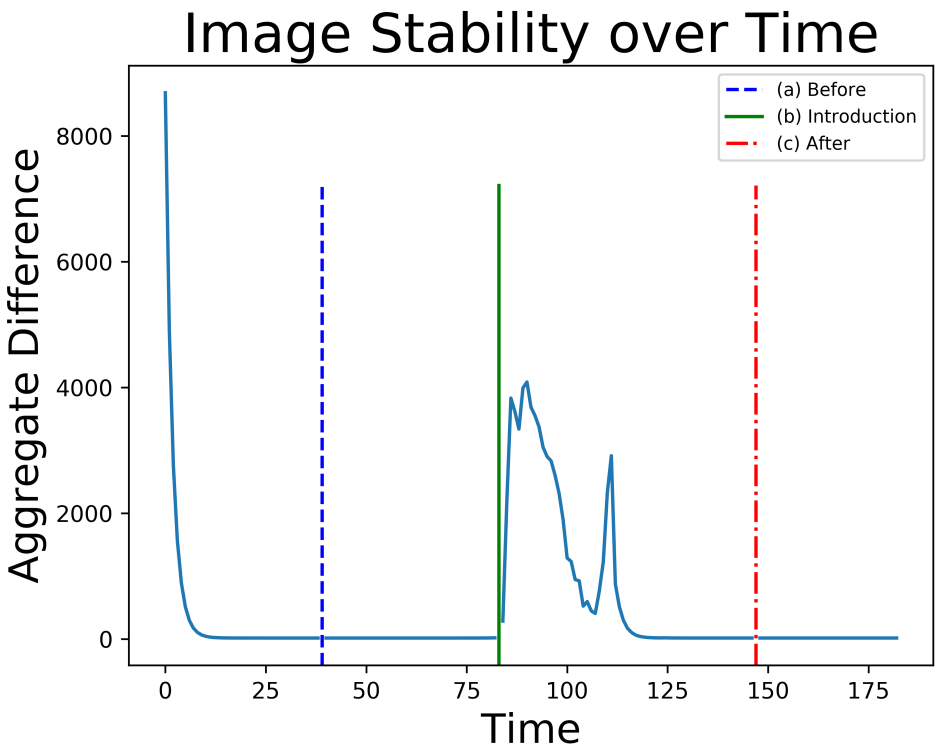
In this section a method for generating a synthetic dataset from a single object useful for training an object detection model is described. The method entails superimposing the object onto a variety of backgrounds to generate a corpus of images with corresponding bounding boxes. Data augmentation strategies are then employed so the model can generalize to conditions not captured in the original image of the object. Generating such a dataset requires an image of the object-of-interest provided by the user along with a bounding box. Requesting that a user with a visual impairment hand-label an image of an object is impractical, so instead a method for automatic object segmentation from a video stream was developed for this purpose.

7.2.1 Automatic Object Segmentation

The data collection process is designed such that a user must simply position a stationary camera, start the process, and place an object into the field of view. Capturing the object works by detecting moments of stability in the video stream and inferring those moments to be frames before and after the introduction of an object. A stability curve and corresponding frame captures are shown in figure 7.2.1. Stability is measured as the aggregate difference between the current frame and a moving average of the previos 10 frames.

Soon after the stream has begun, a frame is captured once stability in the frame is detected (figure 7.3a). A user then places on object in front of the camera, prompting large differences between frames. Another frame is captured after stability returns (7.3b). This process is plotted against frame stability over time in figure 7.2.1.

After the frames are captured, the object is segmented by the system. The resulting "before" and "after" images are diffed(see figure 7.3c) and thresholded on lightness



(a)

Figure 7.2: Frame stability over time. (a) corresponds to the first frame captured (figure 7.3a), (b) indicates when the introduction of the object was detected (c) indicates when the last "after" image was captured (figure 7.3b).

(see figure 7.3d) to create an image mask. To determine a value for a threshold, the algorithm performs a sweep of different threshold, starting with a threshold resulting in 0% of the pixels are thresholded and moving up until 97% of the pixels are thresholded. The threshold is chosen at the point between which the largest change in thresholded pixels is obtained. The thresholded image produces a binary mask of pixels that have changed between frames. From this mask, bounding boxes are infered by constricting the image iteratively from the four directions. The sides of the images constrict the image until 95% of the binary mask remains, creating a

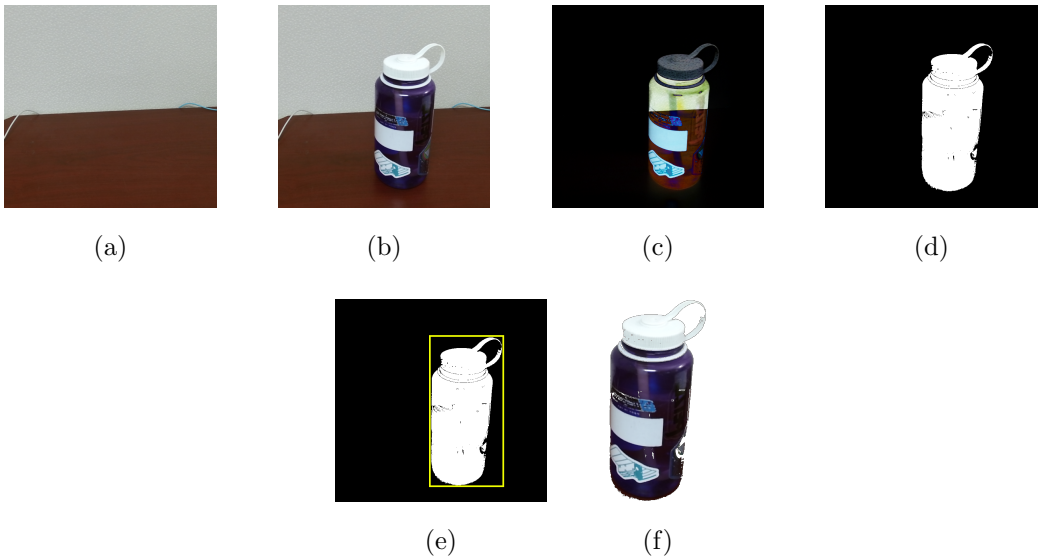


Figure 7.3: (a) Image before object is introduced (b) Image after object is introduced (c) Difference image between "before" and "after" images (d) Thresholded difference image (e) Bounding box applied to image mask by constricting all sides (f) Object after automatic segmentation and cropping has been applied

bounding box containing only the object that was introduced, see figure 7.3e. The image is cropped along the bounding box and the binary mask is applied to the "after" image to generate a masked object that can be artificially placed in other images to generate a synthetic dataset (see figure 7.2.1). For a visual demonstration of this method, see the following video example link: *bottle, ruler, wallet, shoe*. This image, corresponding to figure 7.3f, will be refered to as the **masked object image**.

7.2.2 Data Augmentation

Data augmentation strategies have been shown to significantly increase the efficacy of object detectors trained on limited data, Zoph *et al.* (2019). Due to the minimal effort mandate of this approach, leveraging the limited data that is available

(a) (b)

Figure 7.4: (a) Original image of a counter (b) Masked object image placed onto the original image to create a synthetic data sample.

is essential. To generate the synthetic dataset on which to train an object detection model, a variety of data augmentation strategies are applied to the masked object image such as rotation and perspective transforms (shown in figure 7.2.2) as well as color space tranforms (shown in figure 7.2.2). The transformed images are then placed in random locations and at random scales in a corpus of "background images" which undergo randomized data augmentation as well. The bounding box labels of the object are inferred from the position of the masked object image in the background image after all transformations have been applied and the synthetic images and labels together make up the synthetic training set. A batch of a synthetic dataset can be seen in section D.

7.2.3 Training an Object Detection Model

To validate this approach, an object detection model was trained against the synthetic dataset. The model chosen was YOLOv3, Redmon and Farhadi (2018), based on the original YOLO model for both its respectable performance in both speed and mAP. A preliminary dataset with ground truth was collected containing two "objects of interest", a water bottle and a set of keys, to validate this approach.

(a) (b) (c)

Figure 7.5: Coordinate space augmentations (a) Original "background image" (b) Rotation (c) Perspective

(a) (b) (c)

Figure 7.6: Color space augmentations (a) Hue (b) Saturation (c) Lightness

7.3 Results and Discussion

The following results are derived after training YOLOv3 with 1SODDA for 1 hour (2000 training steps). Table 7.1 compares the performance of several few-shot learning methods on a single training sample to 1SODDA. True Positives for Mean Average Precision (mAP) calculations are detections with an Intersection over Union (IoU) greater than 0.5. The methods compared are YOLO with joint, fine-tuning, and fine-tuning of the full model, Redmon *et al.* (2016), Low-Shot Transfer Detector (LSTD), Chen *et al.* (2018), applied to YOLO, and Few-Shot Object Detection via Feature Reweighting FSOD(RW), Kang *et al.* (2020). It is important to note that these results between this method and other few-shot learning methods are not

Method	**Novel Set 1**	**Novel Set 2**	**Novel Set 3**
YOLO-joint	0.0	0.0	1.8
YOLO-ft	3.2	8.2	8.1
YOLO-ft-full	6.6	12.5	13.0
LSTD(YOLO)	6.9	9.9	10.9
LSTD(YOLO)-full	8.2	11.4	12.6
FSOD(RW)	14.8	15.7	21.3
1SODDA (Ours)	**45.1 ± 1.04**		

Table 7.1: Comparison of performance, mAP (%), between few-shot object detection methods for single-shot learning on the COCO dataset. Originally reported by Kang *et al.* (2020). Performance of 1SODDA is reported with a 95% confidence interval (n=200).

completely comparable, due to the novel data collection method 1SODDA employs that gives a rough segmentation mask instead of purely bounding boxes. This is a richer labeling scheme but requires less effort that traditional hand labelling using the Automatic Object Segmentation approach. Even so, 1SODDA performs very well in comparison to other few-shot methods, outperforming all of them by tens of mAP points. Although not a true comparison because different training and testing sets were used, the performance of 1SODDA is even comparable to high framerate non-few-shot object detection methods such as the original YOLOv3 and SSD300, Liu *et al.* (2016).

For a video demonstration of the trained model's performance, see the following link: *video demo*. A qualitative comparison of 1SODDA with YOLOv3 trained on COCO with only the "bottle" class enabled, is shown in figure 7.7. The proposed

method detects the novel object in a variety of scenarios while YOLOv3+COCO often misses the object altogether. While the YOLOv3+COCO model impressively generalizes the "bottle" class, the non-specificity makes the outputs unsable for a person-specific application. Failure cases of 1SODDA are shown in figure 7.3.

In conclusion, an approach for training any object detection model to detect user-specified objects with minimal effort called 1SODDA was introduced. Additionally, a hand labelled dataset to validate this approach was collected for two user-specified objects, a water bottle and keys. The performance of this one-shot learning approach on the validation dataset is especially promising, significantly outperforming other few-shot methods (in one-shot mode), although not on the same validation data. Future work includes collecting a larger (with more "objects of interest") dataset for a more direct comparison between few-shot methods and standard object detection methods on the COCO object detection standard. Direct comparisons cannot be made without a large corpus of same-instance object data. Additionally, improving the data augmentation process using GANs to photorealistically blend object and background images may also prove useful. Combining 1SODDA with other few-shot methods, such as two-step fine-tuning may also improve performance. Moreover, integrating this method with sensory substitution methods such as Foveated Haptic Gaze to facilitate visually finding objects for people with visual impairments to generalize FHG to real-world applications.

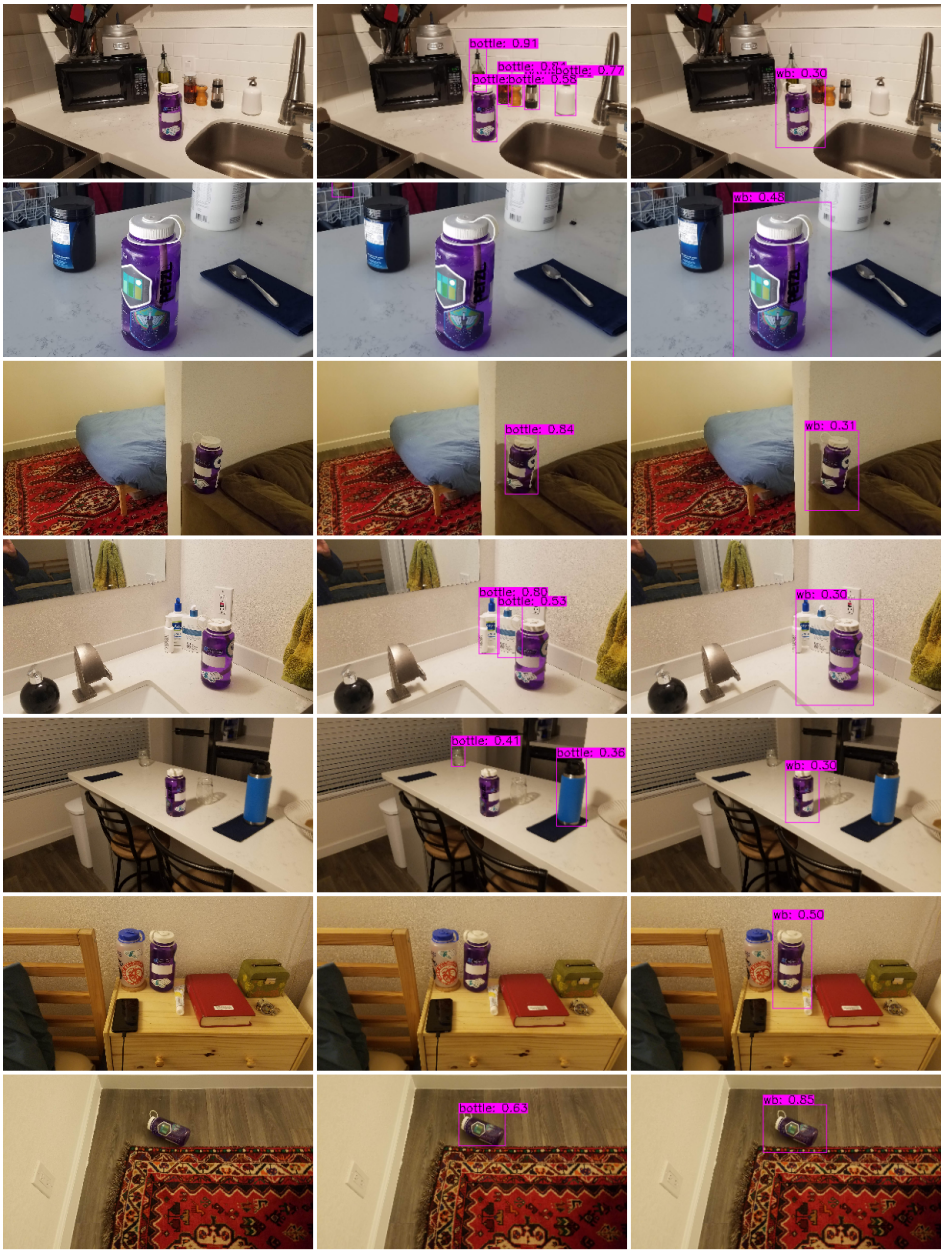


Figure 7.7: Left: Original Image, Middle: Pretrained YOLO model with class "bottle" of COCO dataset, Right: 1SODDA

(a) (b) (c)

Figure 7.8: (a) Missattribution of class failure (b) Missattribution of class failure (c) Failure to detect bottle.

Chapter 8

CONCLUSION AND FUTURE DIRECTIONS

The Person-Centered object detection method, paired with Foveated Haptic Gaze, has the potential to reap the benefits of the advances in Computer Vision in the last decade for use in Sensory Substitution Devices. With self-driving cars and automated medical diagnostics on the horizon, it is unthinkable that the methods behind these technologies are not yet viable for enriching the lives of people with visual impairments. Attempts so far have just grazed the surface of possiblity, showcasing impressive uses of Computer Vision, but their applicability to real-world scenarios is hindered by their rigid nature, lack of person-centeredness, and inefficiencies in real-time interactive environments.

To address these limitations, an intuitive and biologically inspired method for interacting with dynamic visual environments (both virtual and real-world) called Personal Foveated Haptic Gaze was introduced. This approach is composed of two primary components, Foveated Haptic Gaze and a 1-shot object detection method called 1SODDA. Foveated Haptic Gaze and the hardware components used to make it viable such as the Low Resolution Haptic Interface were validated in interactive scenarios via user studies featuring interactive games, both 2-dimensional and 3-dimensional. These methods are predicated on the fact that the visual environment can be parsed in a meaningful way to detect objects of interest in the scene. For these methods to generalize to the real-world, one without ground truth labels, the minimal effort and person-centered approach to realizing object detection models, 1SODDA, was introduced. This method allows a user to employ off-the-shelf and state-of-the-art object detectors to detect specific objects of interest. While direct

comparisons between 1SODDA and few-shot object detectors is not possible without a large corpus of single-instance bounding box annotated images, the performance of 1SODDA on a small dataset collected for this purpose is especially promising. The popular YOLOv3 model was trained with 1SODDA, achieving significantly higher mAP50 scores than other modern few-shot methods. Paired with Foveated Haptic Gaze this becomes Person Centered Foveated Haptic Gaze, and has the potential to be used by people with visual impairments to find and interact with objects of their choosing for which there are no labelled datasets available for, bringing forth a new generation of Haptic Sensory Substitution Devices for Vision which adapt to the needs of the user and harness modern techniques in Artificial Intelligence.

Future work includes performing a large scale user study to assess the effectiveness of PCFHG in a non-virtual setting. Such a study would involve participants with blindness as well as sighted participants, whereby their performance in finding their own objects in a visually crowded space are compared with and without access to PCFHG. Future work to improve the experience also involves unsupervised and semi-supervised approaches to the learning component. Given ego-centric video, learning the visual characteristics of the objects a person uses most may be done in a completely unsupervised manner. Hand detection techniques and datasets such as EgoHands, Bambach *et al.* (2015), could be leveraged to support these efforts. The unsupervised approach could reduce the effort required by the user even more, removing the need for a user to identify and submit to the system objects of interest. An example of this method is illustrated in section C. Furthermore, to increase the intuitiveness of the system, haptic codes for the objects could be proposed by the system and assigned to objects that are commonly used. These methods, coupled with the advances in mobile computing, will hopefully endow users with visual impairments with more independence and agency in the visual world.

REFERENCES

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems", URL `http://arxiv.org/abs/1603.04467` (2016).

Abboud, S., S. Hanassy, S. Levy-Tzedek, S. Maidenbaum and A. Amedi, "EyeMusic: Introducing a 'visual' colorful experience for the blind using auditory sensory substitution", Restorative Neurology and Neuroscience **32**, 2, 247–257 (2014).

Alex Colgan, "How Does the Leap Motion Controller Work?", URL `http://blog.leapmotion.com/hardware-to-software-how-does-the-leap-motion-controller-work/` (2014).

Allamanis, M., M. Brockschmidt and M. Khademi, "Learning to Represent Programs with Graphs", ICLR URL `http://arxiv.org/abs/1711.00740` (2018).

Ando, B., S. Baglio, V. Marletta and A. Valastro, "A Haptic Solution to Assist Visually Impaired in Mobility Tasks", IEEE Transactions on Human-Machine Systems **45**, 5, 641–646 (2015).

Apple, "Timecrest: The Door", https://apps.apple.com/za/app/timecrest-the-door/id1027546326, URL `https://apps.apple.com/za/app/timecrest-the-door/id1027546326` (2015).

Baccus, S. A. and M. Meister, "Fast and slow contrast adaptation in retinal circuitry", Neuron **36**, 5, 909–919 (2002).

Bach-Y-Rita, P., C. C. Collins, F. A. Saunders, B. White and L. Scadden, "Vision substitution by tactile image projection", Nature **221**, 5184, 963–964 (1969).

Bach-y Rita, P., Y. Danilov, M. Tyler and R. J. Grimm, "Late human brain plasticity: vestibular substitution with a tongue BrainPort human-machine interface", Plasticidad y Restauracion Neurologica **4**, 1-2, 31–34, URL `http://www.medigraphic.com/pdfs/plasticidad/prn-2005/prn051{_}2f.pdf{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/15194608{%}0Ahttp://doi.wiley.com/10.1196/annals.1305.006` (2005).

Bala, S., T. McDaniel and S. Panchanathan, "Visual-to-tactile mapping of facial movements for enriched social interactions", 2014 IEEE International Symposium on Haptic, Audio and Visual Environments and Games, HAVE 2014 - Proceedings pp. 82–87 (2014).

Bambach, S., S. Lee, D. J. Crandall and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions", Proceedings of the IEEE International Conference on Computer Vision **2015 Inter**, 1949–1957 (2015).

Barry, N., "Papa Sangre - The Videogame With No Video", URL `https://www.wired.com/2011/01/papa-sangre-the-videogame-with-no-video/` (2011).

Barstow, C. and C. Rerucha, "Evaluation of short and tall stature in children", American Family Physician **92**, 1, 43–50 (2015).

Bengio, Y., "Continuous control with deep reinforcement learning", Foundations and Trends® in Machine Learning **2**, 1, 1–127, URL `https://arxiv.org/pdf/1509.02971.pdf` (2009).

Benjamin, J. M., "The laser cane.", Bulletin of prosthetics research pp. 443–50, URL `http://www.ncbi.nlm.nih.gov/pubmed/4462934` (1974).

Biswas, S. K. and P. Milanfar, "Laplacian object: One-shot object detection by locality preserving projection", 2014 IEEE International Conference on Image Processing, ICIP 2014 pp. 4062–4066 (2014).

Bliss, J. C., M. H. Katcher, C. H. Rogers and R. P. Shepard, "Optical-to-Tactile Image Conversion for the Blind", IEEE Transactions on Man-Machine Systems **11**, 1, 58–65 (1970).

Bourne, R. R., S. R. Flaxman, T. Braithwaite, M. V. Cicinelli, A. Das, J. B. Jonas, J. Keeffe, J. Kempen, J. Leasher, H. Limburg, K. Naidoo, K. Pesudovs, S. Resnikoff, A. Silvester, G. A. Stevens, N. Tahhan, T. Wong, H. R. Taylor, P. Ackland, A. Arditi, Y. Barkana, B. Bozkurt, R. Wormald, A. Bron, D. Budenz, F. Cai, R. Casson, U. Chakravarthy, N. Congdon, T. Peto, J. Choi, R. Dana, M. Palaiou, R. Dandona, L. Dandona, T. Shen, I. Dekaris, M. Del Monte, J. Deva, L. Dreer, M. Frazier, L. Ellwein, J. Hejtmancik, K. Frick, D. Friedman, J. Javitt, B. Munoz, H. Quigley, P. Ramulu, A. Robin, J. Tielsch, S. West, J. Furtado, H. Gao, G. Gazzard, R. George, S. Gichuhi, V. Gonzalez, B. Hammond, M. E. Hartnett, M. He, F. Hirai, J. Huang, A. Ingram, C. Joslin, R. Khanna, D. Stambolian, M. Khairallah, J. Kim, G. Lambrou, V. C. Lansingh, P. Lanzetta, J. Lim, K. Mansouri, A. Mathew, A. Morse, D. Musch, V. Nangia, M. Battaglia, F. Yaacov, M. Raju, L. Rossetti, J. Saaddine, M. Sandar, J. Serle, R. Shetty, P. Sieving, J. C. Silva, R. S. Sitorus, J. Tejedor, M. Tsilimbaris, J. van Meurs, R. Varma, G. Virgili, J. Volmink, Y. Xing, N. L. Wang, P. Wiedemann and Y. Zheng, "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis", The Lancet Global Health **5**, 9, e888–e897, URL `http://linkinghub.elsevier.com/retrieve/pii/S2214109X17302930` (2017).

Brockman, G., V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang and W. Zaremba, "OpenAI Gym", URL `http://arxiv.org/abs/1606.01540` (2016).

Bucchieri, V., "Apparatus and method for presenting and controllably scrolling Braille text", URL `https://patents.google.com/patent/US8382480B2/en` (2013).

Capelle, C., C. Trullemans, P. Arno and C. Veraart, "A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution", IEEE Transactions on Biomedical Engineering **45**, 10, 1279–1293 (1998).

Caspi, A., J. D. Dorn, K. H. McClure, M. S. Humayun, R. J. Greenberg and M. J. McMahon, "Feasibility study of a retinal prosthesis:Spatial vision with a 16-electrode implant", Archives of Ophthalmology **127**, 4, 398–401 (2009).

Cassinelli, A., E. Sampaio, S. B. Joffily, H. R. Lima and B. P. Gusmão, "Do blind people move more confidently with the Tactile Radar?", Technology and Disability **26**, 2-3, 161–170 (2014).

Chabris, C. F., A. Weinberger, M. Fontaine and D. J. Simons, "You do not talk about fight club if you do not notice fight club: Inattentional blindness for a simulated real-world assault", i-Perception **2**, 2, 150–153 (2011).

Chai, C., B. Lau and Z. Pan, "Hungry CatA Serious Game for Conveying Spatial Information to the Visually Impaired", Multimodal Technologies and Interaction **3**, 1, 12, URL `http://www.mdpi.com/2414-4088/3/1/12` (2019).

Chebat, D. R., S. Maidenbaum and A. Amedi, "Navigation using sensory substitution in real and virtual mazes", PLoS ONE **10**, 6, 1–18 (2015).

Chen, H., Y. Wang, G. Wang and Y. Qiao, "LSTD: A low-shot transfer detector for object detection", 32nd AAAI Conference on Artificial Intelligence, AAAI 2018 pp. 2836–2843 (2018).

Clary, P., "Lookout: an app to help blind and visually impaired people learn about their surroundings", https://www.blog.google/outreach-initiatives/accessibility/lookout-app-help-blind-and-visually-impaired-people-learn-about-their-surroundings/, URL `https://www.blog.google/outreach-initiatives/accessibility/lookout-app-help-blind-and-visually-impaired-people-learn-about-their-surroundings/` (2018).

Colwell, C., H. Petrie, D. Kornbrot, A. Hardwick and S. Furner, "Haptic virtual reality for blind computer users", in "Assets '98 Proceedings of the third international ACM conference on Assistive technologies", pp. 92–99 (Marina del Rey, California, USA, 1998), URL `https://dl.acm.org/citation.cfm?id=274515`.

Connors, E. C., L. A. Yazzolino, J. Sánchez and L. B. Merabet, "Development of an Audio-based Virtual Gaming Environment to Assist with Navigation Skills in the Blind", Journal of Visualized Experiments , 73 (2013).

De Felice, F., F. Renna, G. Attolico and A. Distante, "A haptic/acoustic application to allow blind the access to spatial information", Proceedings - Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, World Haptics 2007 pp. 310–315 (2007).

DMNagel, "Timecrest: The Door", https://www.applevis.com/apps/ios/games/timecrest-door, URL https://www.applevis.com/apps/ios/games/timecrest-door (2017a).

DMNagel, "Timecrest: The Door", URL https://www.applevis.com/apps/ios/games/timecrest-door (2017b).

Dowino, "A Blind Legend", https://play.google.com/store/apps/details?id=com.dowino.ABlindLegend (2019).

Duan, Y., J. Schulman, X. Chen, P. Bartlett, I. Sutskever and P. Abbeel, "RLˆ2: Fast Reinforcement Learning Via Slow Reinforcement Learning", arXiv pp. 1–14, URL https://arxiv.org/pdf/1611.02779.pdf (2016).

Duarte, B., T. McDaniel, A. Chowdhury, S. Gill and S. Panchanathan, "HaptWrap: Augmenting Non-Visual Travel via Visual-to-Tactile Mapping of Objects in Motion", in "ACM Multimedia Workshop on Multimedia for Accessible Human Computer Interfaces", (Association for Computing Machinery, Nice, France, 2019).

Duthey, B., "Background Paper 6.11 Alzheimer Disease and other Dementias, Update on 2004", World Health Organization , February, 1 – 77, URL http://www.who.int/medicines/areas/priority{_}medicines/BP6{_}11Alzheimer.pdf (2013).

Eagleman, D., "Plenary talks: A vibrotactile sensory substitution device for the deaf and profoundly hearing impaired", in "2014 IEEE Haptics Symposium (HAPTICS)", pp. xvii–xvii (2014), URL http://ieeexplore.ieee.org/document/6775419/.

Ekstrom1, A. D., "Why vision is important to how we navigate", Hippocampus **73**, 4, 389–400 (2015).

Eskildsen, P., A. Morris, C. C. Collins and P. Bach-y Rita, "Simultaneous and successive cutaneous two-point thresholds for vibration", Psychonomic Science **14**, 4, 146–147 (1969).

Fakhri, B., S. Sharma, B. Soni and A. Chowdhury, "A Low Resolution Haptic Interface for Interactive Applications", HCI International pp. 1–6 (2019).

Fallis, A., "'Smart' Cane for the Visually Impaired: Design and Controlled Field Testing of an Affordable Obstacle Detection System", 12th International Conference on Mobility and Transport for Elderly and Disabled Persons **53**, 9, 1689–1699 (2010).

Fan, Q., W. Zhuo, C.-K. Tang and Y.-W. Tai, "Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector", URL http://arxiv.org/abs/1908.01998 (2019).

Freedom Scientific Inc., "Freedom Scientific Braille Displays and Keyboards", http://www.freedomscientific.com/, URL http://www.freedomscientific.com/ (2018).

Geldard, F. A., "Cutaneous coding of optical signals: The optohapt", Perception & Psychophysics **1**, 11, 377–381 (1966).

Geldard, F. A. and C. E. Sherrick, "The cutaneous "rabbit": A perceptual illusion", Science **178**, 4057, 178–179 (1972).

GHARIEB, W. and G. NAGIB, "Smart Cane for Blinds", Proc. 9th Int. Conf. on AI Applications , August, 253–262, URL c:{%}5CUsers{%}5Cjessica{%}5CBIBLIOTECA{%}5Cdesign{%}5CMetodologiaexperimental-GuiBonsiepe.pdf (2015).

Girshick, R., "Fast R-CNN", Proceedings of the IEEE International Conference on Computer Vision **2015 Inter**, 1440–1448 (2015).

Girshick, R., J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **1**, 580–587, URL http://openaccess.thecvf.com/content{_}cvpr{_}2014/papers/Girshick{_}Rich{_}Feature{_}Hierarchies{_}2014{_}CVPR{_}paper.pdf (2014).

Graves, A., G. Wayne and I. Danihelka, "Neural Turing Machines", Arxiv pp. 1–26, URL http://arxiv.org/abs/1410.5401 (2014).

Graves, A., G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. Gómez Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, A. Puigdomènech Badia, K. Moritz Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, D. Hassabis, A. P. Badia, K. Moritz Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, D. Hassabis, A. Puigdomènech Badia, K. Moritz Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu and D. Hassabis, "Hybrid computing using a neural network with dynamic external memory", URL https://www.nature.com/nature/journal/v538/n7626/pdf/nature20101.pdfhttp://dx.doi.org/10.1038/nature20101http://www.nature.com/nature/journal/v538/n7626/full/nature20101.html (2016).

Gulcehre, C., S. Chandar and Y. Bengio, "Memory Augmented Neural Networks with Wormhole Connections", Arxiv pp. 1–27, URL http://arxiv.org/abs/1701.08718 (2017).

Gupta, S., J. Davidson, S. Levine, R. Sukthankar and J. Malik, "Cognitive Mapping and Planning for Visual Navigation", CVPR URL https://arxiv.org/pdf/1702.03920.pdf (2017).

He, L., Z. Wan, L. Findlater and J. E. Froehlich, "TacTILE: A Preliminary Toolchain for Creating Accessible Graphics with 3D-Printed Overlays and Auditory Annotations", Proc. 19th Int. ACM SIGACCESS Conf. Comput. Access. pp. 397–398, URL http://doi.acm.org/10.1145/3132525.3134818 (2017).

Heess, N., J. J. Hunt, T. P. Lillicrap and D. Silver, "Memory-based control with recurrent neural networks", URL http://arxiv.org/abs/1512.04455 (2015).

Heyes, A. D., "The Sonic Pathfinder: A New Electronic Travel Aid.", Journal of Visual Impairment and Blindness **78**, 5, 200–2, URL http://eric.ed.gov/?id=EJ301318 (1983).

Hochreiter, S. and J. Urgen Schmidhuber, "Long Short-Term Memory", Neural Computation **9**, 8, 1735–1780, URL http://www7.informatik.tu-muenchen.de/{~}hochreit{%}5Cnhttp://www.idsia.ch/{~}juergen (1997).

Homer Jacobson, "The Informational Capacity of the Human Ear", Science **112**, 2901, 143–144, URL http://science.sciencemag.org/content/112/2901/143 (1950).

Hudspeth, A.J.; Schwartz, James; Siegelbaum, Steven; Kandel, Eric; Jessell, T., *Principles of Neural Science, Fifth Edition*, no. 5th (2012).

Ikei, Y., K. Wakamatsu and S. Fukuda, "Vibratory tactile display of image-based textures", IEEE Computer Graphics and Applications **17**, 6, 53–61 (1997).

Inc, S. C., "Timecrest: The Door", URL https://apps.apple.com/za/app/timecrest-the-door/id1027546326 (2015).

Jacobson, H., "The informational capacity of the human eye", Science **113**, 2933, 292–293 (1951).

Jang, E., S. Gu and B. Poole, "Categorical Reparameterization with Gumbel-Softmax", International Conference on Learning Representations pp. 1–13, URL http://arxiv.org/abs/1611.01144 (2017).

Jansson, G., H. Petrie, C. Colwell and D. Kornbrot, "Haptic virtual environments for blind people: Exploratory experiments with two devices", The International Journal of Virtual Reality **3**, 4, 8–17, URL https://pdfs.semanticscholar.org/348e/45107167a0325051e60c883c153572a127e4.pdf{%}0Ahttp://www.ijvr.org/issues/pre/4-1/2.pdf (1999).

Jernigan Institute, "The Braille Literacy Crisis in America, Facing the Truth, Reversing the Trend, Empowering the Blind", Tech. rep., National Federation of the Blind, Baltimore Maryland (2009).

Johnson, M., K. Hofmann, T. Hutton and D. Bignell, "The malmo platform for artificial intelligence experimentation", IJCAI International Joint Conference on Artificial Intelligence **2016-Janua**, 4246–4247, URL https://www.ijcai.org/Proceedings/16/Papers/643.pdf (2016).

Jones, L. A. and K. Ray, "Localization and pattern recognition with tactile displays", Symposium on Haptics Interfaces for Virtual Environment and Teleoperator Systems 2008 - Proceedings, Haptics pp. 33–39 (2008).

Jones, L. A. and N. B. Sarter, "Tactile Displays: Guidance for Their Design and Application", Human Factors: The Journal of the Human Factors and Ergonomics Society **50**, 1, 90–111, URL http://journals.sagepub.com/doi/10.1518/001872008X250638 (2008).

Kang, B., Z. Liu, X. Wang, F. Yu, J. Feng and T. Darrell, "Few-Shot Object Detection via Feature Reweighting", pp. 8419–8428 (2020).

Karlinsky, L., J. Shtok, S. Harary, E. Schwartz, A. Aides, R. Feris, R. Giryes and A. M. Bronstein, "Repmet: Representative-based metric learning for classification and few-shot object detection", in "Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition", vol. 2019-June, pp. 5192–5201 (2019).

Karpathy, A. and F. F. Li, "Deep visual-semantic alignments for generating image descriptions", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **39**, 4, 3128–3137, URL https://arxiv.org/pdf/1412.2306v2.pdf (2015).

Karyn, G., M. Gina, M. Alessandra, C. Robert, B. Peter and C. Graeme, "A comparison of Tactaid II+ and Tactaid 7 use by adults with a profound hearing impairment", Ear and Hearing **20**, 6, 471–482, URL http://www.scopus.com/inward/record.url?eid=2-s2.0-0033436316{&}partnerID=40{&}md5=1058dec9323d2a378c6fcb685db9acc5 (1999).

Kay, L., "A sonar aid to enhance spatial perception of the blind : engineering design and evaluation", The Radio and Electronic Engineer **44**, 11, 605–627 (1974).

Kempka, M., M. Wydmuch, G. Runc, J. Toczek and W. Jaskowski, "ViZDoom: A Doom-based AI research platform for visual reinforcement learning", IEEE Conference on Computatonal Intelligence and Games, CIG URL https://arxiv.org/pdf/1605.02097.pdf (2017).

Koch, G., R. Zemel and R. Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition", in "ICML Workshop", (2015), URL https://www.cs.cmu.edu/{~}rsalakhu/papers/oneshot1.pdf.

König, H., J. Schneider and T. Strothotte, "Haptic Exploration of Virtual Buildings Using Non-Realistic Haptic Rendering", Society pp. 377–384 (2000).

König, H., J. Schneider and T. Strothotte, "Orientation and Navigation in Virtual Haptic-Only Environments", in "Proceedings User Guidance in Virtual Environments", edited by V. Paelke and S. Volbracht, pp. 123–136 (Shaker Verlag, Aachen, Germany, 2001).

Krantz, J. H., "The Stimulus and Anatomy of the Visual System", Experiencing Sensation and Perception pp. 3.1 – 3.36, URL https://psych.hanover.edu/classes/sensation/chapters/Chapter3.pdf (2012).

Krishna, S., S. Bala, T. McDaniel, S. McGuire and S. Panchanathan, "VibroGlove: an assistive technology aid for conveying facial expressions", in "CHI '10 Extended Abstracts on Human Factors in Computing Systems", pp. 3637–3642 (2010), URL http://doi.acm.org/10.1145/1753846.1754031.

Krizhevsky, A., I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Advances In Neural Information Processing Systems pp. 1–9, URL https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (2012).

Kuc, R., "Binaural sonar electronic travel aid provides vibrotactile cues for landmark, reflector motion and surface texture classification", IEEE Transactions on Biomedical Engineering **49**, 10, 1173–1180 (2002).

Lahav, O. and D. Mioduser, "Construction of cognitive maps of unknown spaces using a multi-sensory virtual environment for people who are blind", Computers in Human Behavior **24**, 3, 1139–1155 (2008).

Landau, S. and L. Wells, "Merging Tactile Sensory Input and Audio Data by Means of The Talking Tactile Tablet", Proc. Eurographics'03 **2**, 60, 414–418 (2003).

Lévesque, V., J. Pasquero, V. Hayward and M. Legault, "Display of virtual braille dots by lateral skin deformation: feasibility study", ACM Transactions on Applied Perception **2**, 2, 132–149 (2005).

Li, Z. and N. Snavely, "MegaDepth: Learning Single-View Depth Prediction from Internet Photos", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 2041–2050 (2018).

Lindeman, R. W. and Y. Yanagida, "Empirical studies for effective near-field haptics in virtual environments", Proceedings - IEEE Virtual Reality **2003-Janua**, 287–288 (2003).

Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu and A. C. Berg, "SSD: Single shot multibox detector", in "Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)", vol. 9905 LNCS, pp. 21–37 (2016).

Loomis, J. M., "Tactile letter recognition under different modes of stimulus presentation", Perception & Psychophysics **16**, 2, 401–408 (1974).

Maidenbaum, S., S. Levy-Tzedek, D. R. Chebat and A. Amedi, "Increasing accessibility to the blind of virtual environments, using a virtual mobility aid based on the "EyeCane": Feasibility study", PLoS ONE **8**, 8 (2013).

Maidenbaum, S., S. Levy-Tzedek, R. Namer-Furstenberg, A. Amedi and D. R. Chebat, "The effect of extended sensory range via the eyecane sensory substitution device on the characteristics of visionless virtual navigation", Multisensory Research **27**, 5-6, 379–397 (2014).

Mascetti, S., C. Bernareggi and M. Belotti, "TypeInBraille: A Braille-based Typing Application for Touchscreen Devices", in "The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility", pp. 295–296 (Dundee, Scotland, UK, 2011).

Mazuryk, T., M. Gervautz and K. Smith, "Virtual Reality History, Applications, Technology and Future", Digital Outcasts **63**, ISlE, 92–98, URL `http://www.sciencedirect.com/science/article/pii/B9780124047051000066{%}5Cnhttp://linkinghub.elsevier.com/retrieve/pii/B9780124047051000078{%}5Cn{%}22http://www.cg.tuwien.ac.at/research/publications/1996/mazuryk-1996-VRH/{%}22,{%}7D{%}5Cnhttp://citeseerx.ist.psu.edu` (2013).

McDaniel, T., S. Krishna, V. Balasubramanian, D. Colbry and S. Panchanathan, "Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind", HAVE 2008 - IEEE International Workshop on Haptic Audio Visual Environments and Games Proceedings , October, 13–18 (2008).

McDaniel, T., D. Villanueva, S. Krishna and S. Panchanathan, "MOVeMENT: A framework for systematically mapping vibrotactile stimulations to fundamental body movements", HAVE 2010 - 2010 IEEE International Symposium on Haptic Audio-Visual Environments and Games, Proceedings pp. 13–18 (2010).

McDaniel, T. L., S. Krishna, D. Colbry and S. Panchanathan, "Using tactile rhythm to convey interpersonal distances to individuals who are blind", CHI Extended Abstracts pp. 4669–4674, URL `https://dl.acm.org/citation.cfm?id=1520718` (2009).

Meijer, P. B., "An Experimental System for Auditory Image Representations", IEEE Transactions on Biomedical Engineering **39**, 2, 112–121 (1992).

Menikdiwela, M. P., K. M. Dharmasena and A. M. S. Abeykoon, "Haptic based walking stick for visually impaired people", 2013 International Conference on Circuits, Controls and Communications, CCUBE 2013 pp. 1–6 (2013).

Merzenich, M. M., R. P. Michelson, C. R. Pettit, R. A. Schindler and M. Reid, "Neural Encoding of Sound Sensation Evoked by Electrical Stimulation of the Acoustic Nerve", Annals of Otology, Rhinology & Laryngology **82**, 4, 486–503 (1973).

Metz, C., "Facebook's AI Is Now Automatically Writing Photo Captions", https://www.wired.com/2016/04/facebook-using-ai-write-photo-captions-blind-users/, URL `https://www.wired.com/2016/04/facebook-using-ai-write-photo-captions-blind-users/` (2016).

Metz, R., "BLITAB", https://www.technologyreview.com/s/603336/this-500-tablet-brings-words-to-blind-users-fingertips/, URL https://www.technologyreview.com/s/603336/this-500-tablet-brings-words-to-blind-users-fingertips/ (2017).

Meyer, I. and H. Mikesch, "FEER the Game of Running Blind", http://www.feer.at/index.php/en/home/, URL http://www.feer.at/index.php/en/home/ (2018).

Microsoft, "Seeing AI", https://www.microsoft.com/en-us/ai/seeing-ai, URL https://www.microsoft.com/en-us/ai/seeing-ai (2018).

Miller, I., A. Pather, J. Milbury, L. Hathy, A. O'Day and D. Spence, "Guidelines and Standards for Tactile Graphics, 2010", http://www.brailleauthority.org/tg/web-manual/index.html, URL http://www.brailleauthority.org/tg/web-manual/index.html (2011).

Mirowski, P., M. K. Grimes, M. Malinowski, K. M. Hermann, K. Anderson, D. Teplyashin, K. Simonyan, K. Kavukcuoglu, A. Zisserman and R. Hadsell, "Learning to Navigate in Cities Without a Map", URL http://arxiv.org/abs/1804.00168 (2018).

Mirowski, P., R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu and D. London, "LEARNING TO NAVIGATE IN COMPLEX ENVIRONMENTS", URL https://arxiv.org/pdf/1611.03673.pdf (????).

Mirsky, S., "Playing by Ear", Scientific American **300**, 3, 29–29 (2009).

Miyazaki, M., M. Hirashima and D. Nozaki, "The "Cutaneous Rabbit" Hopping out of the Body", Journal of Neuroscience **30**, 5, 1856–1860, URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3887-09.2010 (2010).

Mnih, V., A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning", arXiv preprint **48**, arXiv:1602.01783v1 [cs.LG], 1–28, URL http://arxiv.org/abs/1602.01783 (2016).

Mnih, V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning", Arxiv URL https://www.cs.toronto.edu/{~}vmnih/docs/dqn.pdf (2013).

Mnih, V., K. Kavukcuoglu, D. Silver, A. a. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg and D. Hassabis, "Human-level control through deep reinforcement learning", Nature **518**, 7540, 529–533, URL http://dx.doi.org/10.1038/nature14236 (2015).

Monacelli, A. M., L. A. Cushman, V. Kavcic and C. J. Duffy, "Spatial disorientation in Alzheimer's disease: The remembrance of things passed", Neurology **61**, 11, 1491–1497, URL https://pdfs.semanticscholar.org/e957/14321d7fb821b421f2897496ccd1d10fed60.pdf (2003).

Moser, M.-B., D. C. Rowland and E. I. Moser, "Place cells, grid cells, and memory.", Cold Spring Harbor perspectives in biology **7**, 2, a021808, URL file:///tmp/mozilla{_}pauli0/ColdSpringHarbPerspectBiol-2015-Moser-.pdf (2015).

Nations, U., "Disability Statistics Compendium", URL http://www.disabilitystatistics.org/ (1990).

Nau, A., M. Bach and C. Fisher, "Clinical Tests of Ultra-Low Vision Used to Evaluate Rudimentary Visual Perceptions Enabled by the BrainPort Vision Device", Translational Vision Science & Technology **2**, 3, 1, URL http://tvst.arvojournals.org/Article.aspx?doi=10.1167/tvst.2.3.1 (2013).

Neisser, U. and R. Becklen, "Selective looking: Attending to visually specified events", Cognitive Psychology **7**, 4, 480–494 (1975).

NFB, "Blindness Statistics — National Federation of the Blind", Tech. rep., National Federation of the Blind, Baltimore, Maryland, URL https://www.nfb.org/resources/blindness-statistics (2017).

Novich, S. D., "Sound-to-Touch Sensory Substitution and Beyond", URL https://scholarship.rice.edu/handle/1911/88379 (2015).

Novich, S. D. and D. M. Eagleman, "Using space and time to encode vibrotactile information: toward an estimate of the skin's achievable throughput", Experimental Brain Research **233**, 10, 2777–2788 (2015).

Oh, J., V. Chockalingam, S. Singh and H. Lee, "Control of Memory, Active Perception, and Action in Minecraft", arXiv:1605.09128 [cs] URL http://arxiv.org/abs/1605.09128{%}5Cnhttp://www.arxiv.org/pdf/1605.09128.pdf (2016).

Panchanathan, S., S. Chakraborty and T. McDaniel, "Social Interaction Assistant: A Person-Centered Approach to Enrich Social Interactions for Individuals with Visual Impairments", IEEE Journal on Selected Topics in Signal Processing **10**, 5, 942–951 (2016).

Pasquero, J., "Survey on communication through touch", McGill Centre for Intelligent Machines **6**, August, 1–28, URL http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:Survey+on+Communication+through+Touch{#}0 (2006).

Pietrzak, T., I. Pecci and B. Martin, "Static and dynamic tactile directional cues experiments with VTPlayer mouse", in "Proceedings of the Eurohaptics conference", pp. 63–68 (Paris, France, 2006).

Redmon, J., S. Divvala, R. Girshick and F. Ali, "(YOLO) You Only Look Once", in "Computer Vision and Pattern Recognition (CVPR)", (Computer Vision Foundation, Las Vegas, Nevada, 2016), URL http://pjreddie.com/yolo/.

Redmon, J. and A. Farhadi, "YOLO v.3", Tech. rep., University of Washington, URL https://pjreddie.com/media/files/papers/YOLOv3.pdf (2018).

Régo, N., "The Game of Running Blind in FEAR", https://coolblindtech.com/the-game-of-running-blind-in-fear/, URL https://coolblindtech.com/the-game-of-running-blind-in-fear/ (2018).

Riener, a. and H. Hartl, "Personal Radar: a self-governed support system to enhance environmental perception", Proceedings of BCS-HCI 2012 pp. 147–156, URL http://dl.acm.org/citation.cfm?id=2377933{%}5Cnpapers://c80d98e4-9a96-4487-8d06-8e1acc780d86/Paper/p15116 (1974).

Rose, D. H. D. H., A. Meyer and C. Hitchcock, *The universally designed classroom : accessible curriculum and digital technologies* (2005), URL https://eric.ed.gov/?id=ED568861.

Rosenthal, J., N. Edwards, D. Villanueva, S. Krishna, T. McDaniel and S. Panchanathan, "Design, implementation, and case study of a pragmatic vibrotactile belt", in "IEEE Transactions on Instrumentation and Measurement", vol. 60, pp. 114–125 (2011).

Sampaio, E., S. Maris and P. Bach-y Rita, "Brain plasticity: 'Visual' acuity of blind persons via the tongue", Brain Research **908**, 2, 204–207 (2001).

SÁNCHEZ, J. and M. LUMBRERAS, "Virtual Environment Interaction Through 3D Audio by Blind Children", CyberPsychology & Behavior **2**, 2, 101–111 (2009).

Sanchez-Gonzalez, A., N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell and P. Battaglia, "Graph networks as learnable physics engines for inference and control", URL http://arxiv.org/abs/1806.01242 (2018).

Santoro, A., S. Bartunov, M. Botvinick, D. Wierstra and T. Lillicrap, "One-shot Learning with Memory-Augmented Neural Networks", arXiv:1605.06065 [cs] URL https://arxiv.org/pdf/1605.06065.pdf (2016).

Schmidt, R. N., F. J. Lisy, T. S. Prince and G. S. Shaw, "Refreshable braille display system", URL https://patents.google.com/patent/US6354839B1/en (1998).

Schulman, J., S. Levine, M. Jordan and P. Abbeel, "Trust Region Policy Optimization", Icml-2015 p. 16, URL http://arxiv.org/abs/1502.0547 (2015).

Semega, J. L., K. R. Fontenot and M. A. Kollar, "Income and poverty in the United States: 2016", Tech. Rep. September, US Census Bureau, URL https://www.census.gov/content/dam/Census/library/publications/2017/demo/P60-259.pdf{%}0Ahttp://www.census.gov/content/dam/Census/library/publications/2016/demo/p60-256.pdf (2017).

Semwal, S., "MoVE: Mobiltiy training in haptic virtual environment", Tech. rep., University of Colorado at Colorado Springs, Colorado Springs, URL https://pdfs.semanticscholar.org/243e/b3d64990f34d0b126b36132acc17e4f50737.pdf{%}0Ahttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.24.6646{&}rep=rep1{&}type=pdf (2001).

Sharkey, P., C. Sik Lanyi, P. Standen and D. o. C. University of Reading. ICDVRAT, "Multisensory virtual environment for supporting blind persons' acquisition of spatial cognitive mapping, orientation, and mobility skills", , 1993, 279 (2002).

Simons, D. J. and C. F. Chabris, "Gorillas in our midst: Sustained inattentional blindness for dynamic events", Perception **28**, 9, 1059–1074 (1999).

Smithmaitrie, P., J. Kanjantoe and P. Tandayya, "Touching force response of the piezoelectric Braille cell", Disability and Rehabilitation: Assistive Technology **3**, 6, 360–365 (2008).

Spence, C., "The skin as a medium for sensory substitution", Multisensory Research **27**, 5-6, 293–312 (2014).

Stageberg, S., "The Device That Refreshes: How to Buy a Braille Display", https://www.afb.org/aw/5/6/14669, URL https://www.afb.org/aw/5/6/14669 (2004).

Sweller, J., P. Ayres and S. Kalyuga, "Amassing information: The information store principle", in "Cognitive Load Theory", (2011).

Szinte, M. and P. Cavanagh, "Apparent Motion from Outside the Visual Field, Retinotopic Cortices May Register Extra-Retinal Positions", PLoS ONE **7**, 10 (2012).

Tan, H. Z. and A. Pentland, "Tactual . Disptays for Wearabte Computing", Personal Technologies pp. 225–230 (1997).

Thurlow, M. L., S. J. Thompson, L. Walz and H. Shin, "Student Perspectives on Using AccommodationsDuring Statewide Testing", Tech. rep., AmericanEducational Research Association, Minneapolis, MN, URL https://files.eric.ed.gov/fulltext/ED474766.pdf (2001).

Troxel, D., "Experiments in Tactile and Visual Reading", IEEE Transactions on Human Factors in Electronics **8**, 4, 261–263, URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp={&}arnumber=1698280 (1967).

Twardon, L., H. Koesling, A. Finke and H. Ritter, "Gaze-contingent audio-visual substitution for the blind and visually impaired", 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops pp. 129–136 (2013).

Type, V. M. and K. Features, "Pico Vibe Model : 310-117 Datasheet", **44**, 0, 1–5 (2016).

Tzovaras, D., K. Moustakas, G. Nikolakis and M. G. Strintzis, "Interactive mixed reality white cane simulation for the training of the blind and the visually impaired", Personal and Ubiquitous Computing **13**, 1, 51–58 (2009).

Van Erp, J. and B. Self, *Tactile Displays for Orientation , Navigation and Communication in Air , Sea and Land Environments*, vol. 323 (2008).

Wall, S. and S. Brewster, "Feeling What You Hear : Tactile Feedback for Navigation of Audio Graphs", in "CHI 2006 Proceedings, Disabilities", pp. 1123–1132 (2006a).

Wall, S. A. and S. Brewster, "Sensory substitution using tactile pin arrays: Human factors, technology and applications", Signal Processing **86**, 12, 3674–3695 (2006b).

Wang, Q., V. Levesque, J. Pasquero and V. Hayward, "A haptic memory game using the STRESS 2 tactile display", in "Proceedings of CHI 2006", p. 271 (2006).

Wang, X., T. E. Huang, T. Darrell, J. E. Gonzalez and F. Yu, "Frustratingly Simple Few-Shot Object Detection", (2020), URL `http://arxiv.org/abs/2003.06957`.

Ward, J. and P. Meijer, "Visual experiences in the blind induced by an auditory sensory substitution device", Consciousness and Cognition **19**, 1, 492–500 (2010).

WebAIM (Web Accessibility In Mind), "Screen Reader User Survey #6", Tech. rep., Utah State University, Logan, Utah, URL `http://webaim.org/projects/screenreadersurvey6/` (2015).

White, B. W., F. A. Saunders, L. Scadden, P. Bach-Y-Rita and C. C. Collins, "Seeing with the skin", Perception & Psychophysics **7**, 1, 23–27 (1970).

WikiHow, "How to Swing a Golf Club", https://www.wikihow.com/Swing-a-Golf-Club, URL `https://www.wikihow.com/Swing-a-Golf-Club` (2019).

Wolbers, T. and M. Hegarty, "What determines our navigational abilities?", Trends in Cognitive Sciences **14**, 3, 138–146, URL `https://ac.els-cdn.com/S1364661310000021/1-s2.0-S1364661310000021-main.pdf?{_}tid=2d79312d-6612-4226-9edc-df132f4f2327{&}acdnat=1529432441{_}c770668cd86ef92d553d968fbd2daa5e` (2010).

Xie, L., S. Wang, A. Markham and N. Trigoni, "Towards Monocular Vision based Obstacle Avoidance through Deep Reinforcement Learning", Robotics: Science and Systems Workshop 2017: New Frontiers for Deep Learning in Robotics URL `http://arxiv.org/abs/1706.09829` (2017).

Yanagida, Y., M. Kakita, R. W. Lindeman, Y. Kume and N. Tetsutani, "Vibrotactile letter reading using a low-resolution tactor array", Proceedings - 12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, HAPTICS pp. 400–406 (2004).

Yang, U., Y. Jang and G. J. Kim, "Designing a VibroTactile Wear for Close Range Interaction for VRbased Motion Training", Icat 2002 pp. 4—-9, URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.103.5793` (2002).

Zhang, J., J. T. Springenberg, J. Boedecker and W. Burgard, "Deep reinforcement learning with successor features for navigation across similar environments", IEEE International Conference on Intelligent Robots and Systems **2017-Septe**, 2371–2378, URL https://arxiv.org/pdf/1612.05533.pdf (2017).

Zhao, Y., C. L. Bennett, H. Benko, E. Cutrell, C. Holz, M. R. Morris and M. Sinclair, "Enabling People with Visual Impairments to Navigate Virtual Reality with a Haptic and Auditory Cane Simulation", Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18 pp. 1–14, URL http://dl.acm.org/citation.cfm?doid=3173574.3173690 (2018).

Zhao, Y., E. Cutrell, C. Holz, M. Morris, E. Ofek and A. Wilson, "SeeingVR: A Set of Tools to Make Virtual Reality More Accessible to People with Low Vision", in "CHI Conference on Human Factors in Computing Systems Proceedings", p. 14 (2019), URL https://doi.org/10.1145/3290605.3300341.

Zhu, J. and J. Yang, "Subpixel eye gaze tracking", Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition pp. 131–136, URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1004144 (2002).

Zoph, B., E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens and Q. V. Le, "Learning Data Augmentation Strategies for Object Detection", URL http://arxiv.org/abs/1906.11172 (2019).

Zou, Z., Z. Shi, Y. Guo and J. Ye, "Object Detection in 20 Years: A Survey", pp. 1–39, URL http://arxiv.org/abs/1905.05055 (2019).

APPENDIX A

LEAP MOTION CONTROLLER

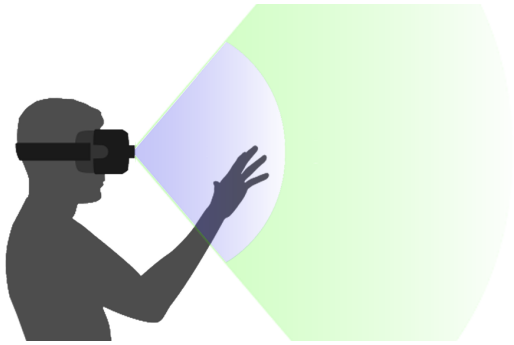


Figure A.1: Leap Mounted on VR Headset

The Leap Motion's field of view (FOV) is 150 deg degrees wide and 120 deg degrees deep. This is not largely different that the 180 deg by 180 deg degree FOV of the human visual system Mazuryk *et al.* (2013). In that field of view, the effective frustrum begins 25mm from the device and ends 600mm from the device for a total of 8 cubic feet of interaction volume Alex Colgan (2014). While this does seem large, it does place some limits on the positioning of the sensor, since a typical adult arm-span is 2.1 inches greater than their height Barstow and Rerucha (2015), resulting in a single-arm span of 3 feet.

Other limitations to consider are bright direct sunlight can affect the Leap Motion (Leap does not gaurantee performance outside), as well as occlusion from other limbs and clothing. Mounting the Leap to a headset (similar to how it is done in VR applications, see figure A.1), would be the best option for ensuring the field of view is unobstructed while maintaining a common coordinate space with the camera (also mounted on the headset). Mounting the Leap on a chest-mount or a belt-mount would reduce the weight of the headset which could be cumbersome, but has the added difficulty of occlusions from clothing (stray shirts), and other objects such as tables. To validate Foveated Haptic Gaze, the Leap Motion was placed on the table in front of the user to avoid the complexities involved in wearables.

APPENDIX B

LRHD VIBROTACTILE PATTERNS

The vibratactile patterns used in the LRHI study Fakhri *et al.* (2019) are shown in figures B.1, B.2 and B.3. Patterns for Phase 1 are static, they do not change over time. Patterns in Phase 2 and 3 though (shown in figure B.2) do evolve over time. The illustration shows how the patterns evolve over time (left to right). The first four patterns (Left-to-Right, Right-to-Left, Top-to-Bottom, and Bottom-to-Top) have 4 states while the last two (Out-to-In and In-to-Out) only have two states. In Phase 2, the patterns lasted a total of 1 second, while in Phase 3 they lasted either 0.6 seconds or 1.4 seconds depending on the speed. In Phase 2 participants were asked to recall what pattern they had experienced and in Phase 3 they were asked what pattern they had experienced as well as how fast the pattern was presented (Slow, Fast).

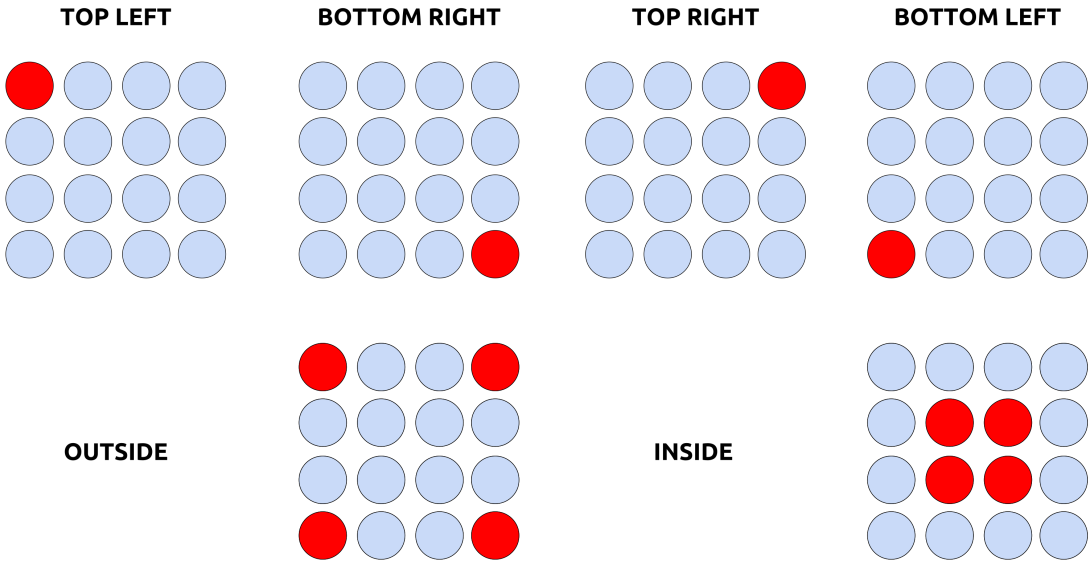


Figure B.1: User Study Patterns for Phase 1: Patterns in this phase do not change over time. They remain static. A red motor in this illustration represents a motors on at full power, while a blue motor represents a motor that is completely inactive (off). Participants were asked to recall what pattern they are experiencing after being subjected to the pattern for 1 second.

In the last phase of the LRHI study (the interactive phase), participants played a two dimensional game depicted in figure B.3. The user plays as a cat that is displayed on the Low Resolution Haptic Display whose goal is to find a mouse. The game is played completely on the haptic display, the user is not given any visual cues. A successful game occurs when the cat finds the mouse (as shown by the arrows). The game is timed: the participant is told to find the mouse as fast as possible. The user controlls the cat using a computer mouse peripheral and the cat moves on the haptic display with respect to the participant's mouse movements.

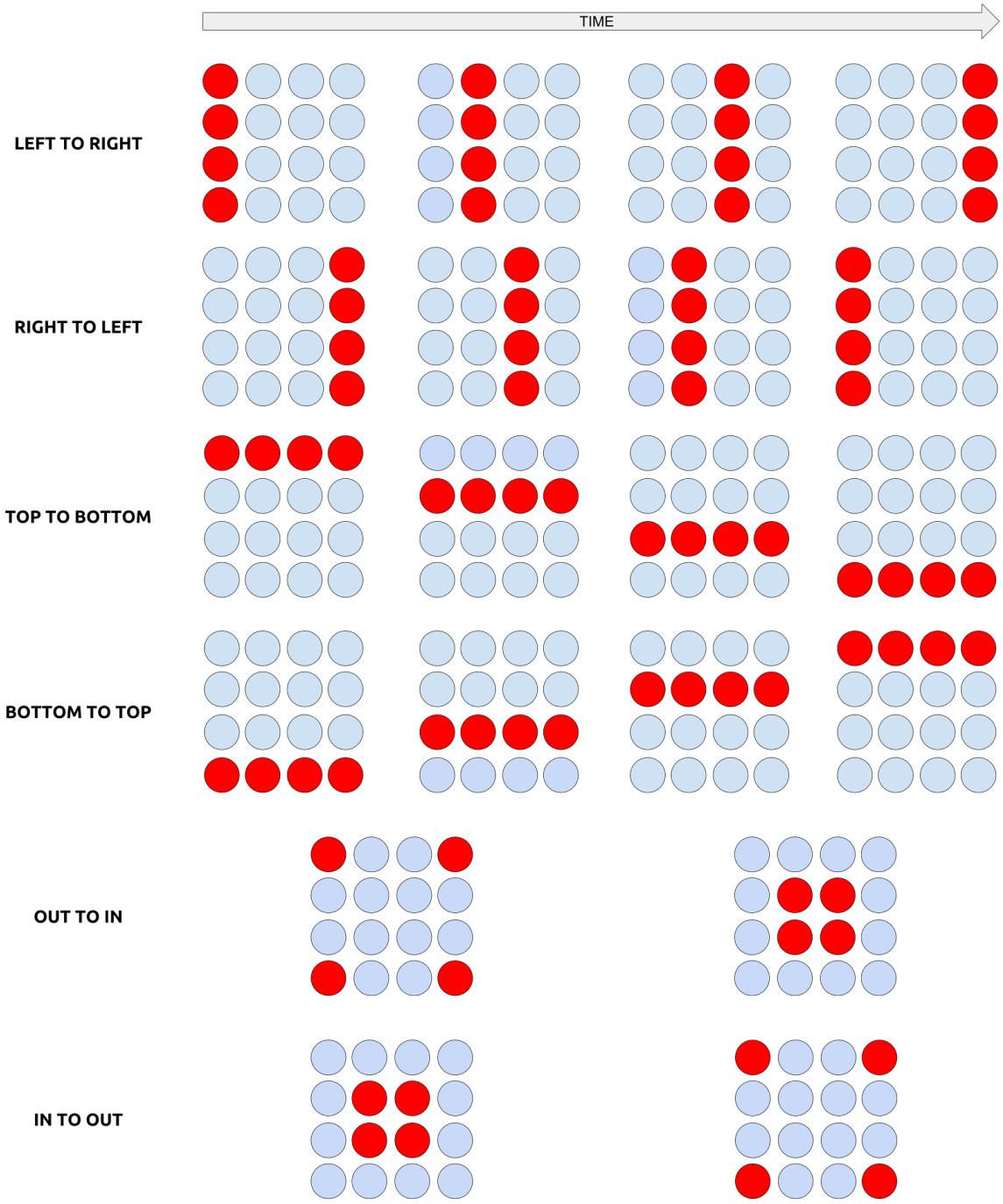


Figure B.2: User Study Patterns for Phases 2 and 3: Patterns in these two phases were dynamic, meaning they evolve over time. A red motor in this illustration represents a motors on at full power, while a blue motor represents a motor that is completely inactive (off).

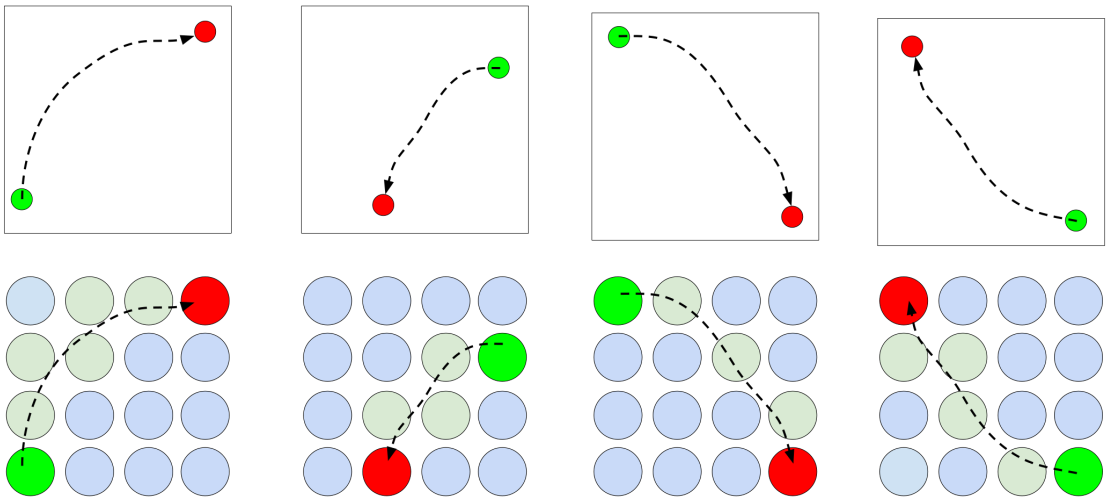


Figure B.3: Illustration of cat-mouse game for The interactive Phase: This is a game where the user plays a cat (green) that has to find a mouse (red). It is depicted visually in the top row. The bottom row depicts the activation of motors on the haptic display. A red motor represents a "pulsating" motor while a green motor is a statically vibrating motor, these represent the mouse and cat respectfully.

APPENDIX C

IN-HAND OBJECT SEGMENTATION

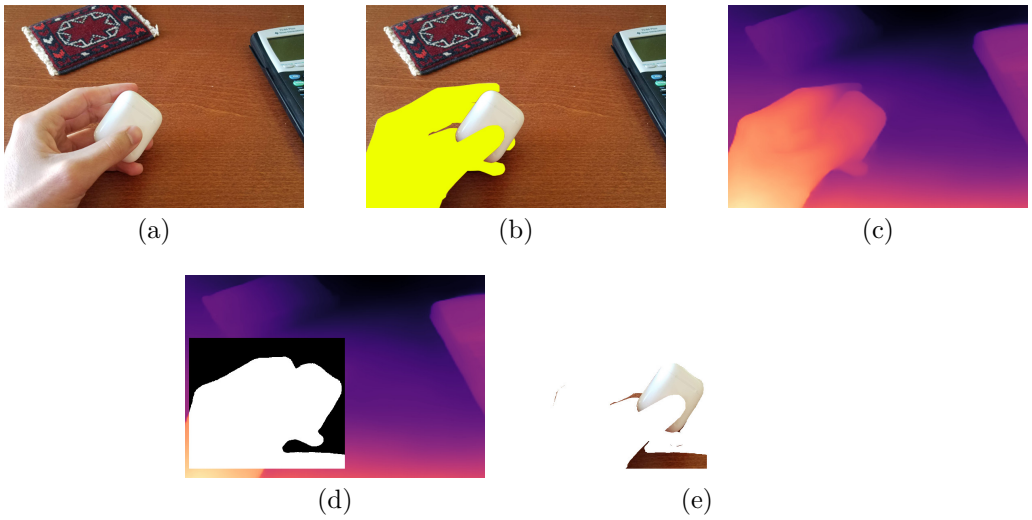


Figure C.1: (a) Original image of a hand holding an object (b) Hand segmentation applied to image (c) Inferred depth-map of image (MegaDepth) (d) Thresholded depth-map in a predefined range over and under the depth at the center of the hand (e) Object masked: *logical and* between hand segmentation and thresholded depth map

Figure C illustrates a naive method for segmenting an object being held from egocentric images. The method requires several components: a hand detection module, hand segmentation module and a depth inference module. The hand detection and segmentation module can be combined, as bounding boxes can be inferred from segmentation masks. Segmenting the objects is performed as follows. The hand detection module produces a bounding box around the hand while a hand segmentation module produces a mask for any hands in the image (see figure C.1b). The depth inference module then produces a depth-map from the image (see figure C.1c). This depth-map is then thresholded using a predefined range with the center of the range defined as the depth at the center of the bounding box of the hand (see figure C.1d). Finally, a segmentation mask for the object in the hand is produced by performing a logical *and* operation between the thresholded depth-map and the hand segmentation mask. This is illustrated in figure C.1e.

APPENDIX D

SYNTHETIC DATASET

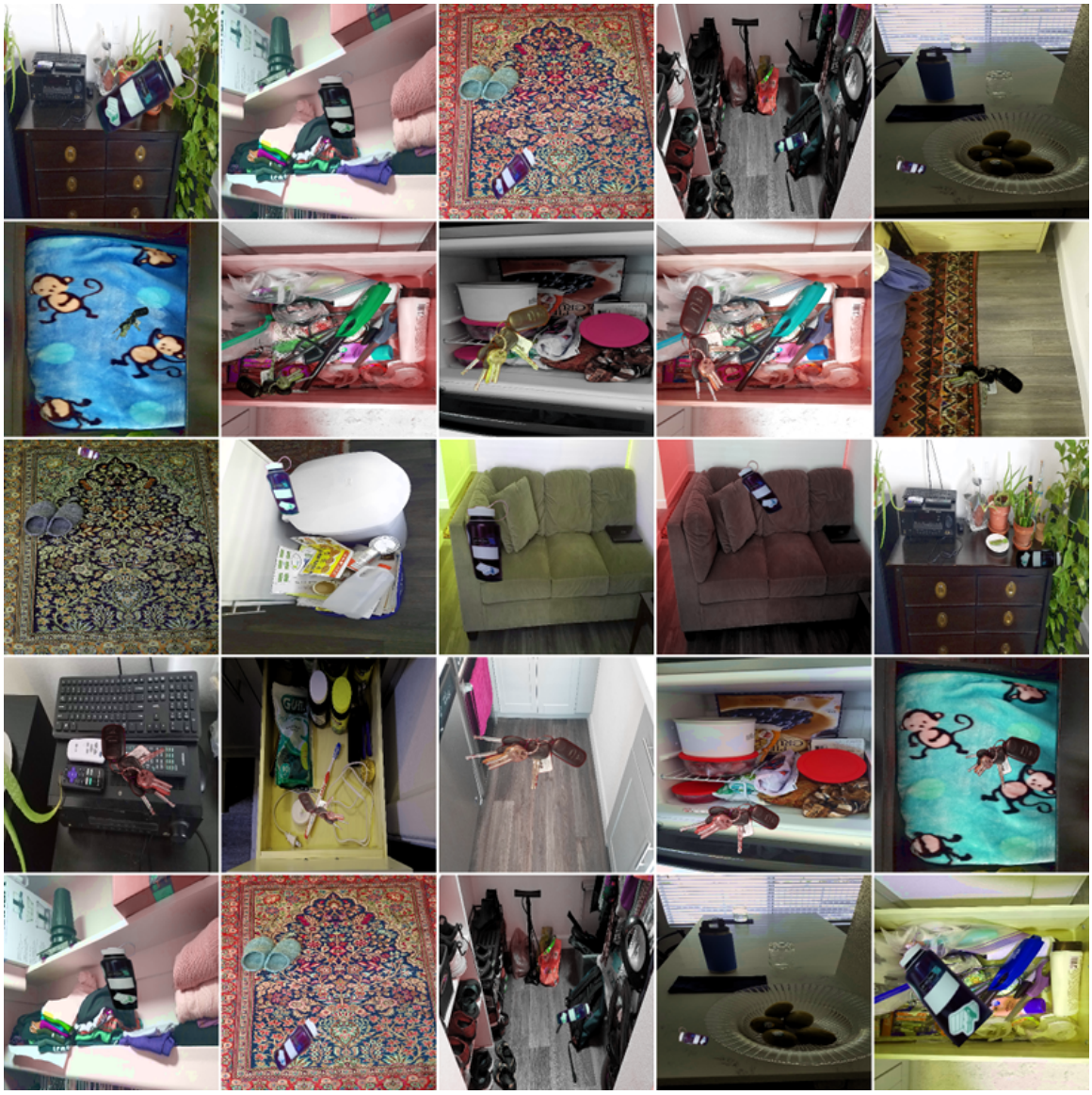


Figure D.1: Samples from the synthetic dataset generated from two objected having undergone the automatic segmentation process. These objects are a water bottle and keys.

APPENDIX E

PERMISSION FROM CO-AUTHORS

Permission for the inclusion of co-authored material in this dissertation was obtained from all co-authors, including: Prof. Sethuraman Panchanathan, Dr. Troy McDaniel, Dr. Hemanth Venkateswara, Dr. Heni Ben Amor, Dr. Zsolt Kira, Aaron Keech, Joel Schlosser, Ethan Brooks, Shashank Sharma, Bhavica Soni, and Abhik Chowdhury.

APPENDIX F

PERMISSION FROM PUBLISHERS FOR RE-PRINT

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Mar 04, 2020

This Agreement between Bijan Fakhri ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4782260426825
License date	Mar 04, 2020
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Haptics for Sensory Substitution
Licensed Content Author	Bijan Fakhri, Sethuraman Panchanathan
Licensed Content Date	Jan 1, 2020
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	electronic

Print This Page

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Mar 04, 2020

This Agreement between Bijan Fakhri ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4782220625111
License date	Mar 04, 2020
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Deep Reinforcement Learning Methods for Navigational Aids
Licensed Content Author	Bijan Fakhri, Aaron Keech, Joel Schlosser et al
Licensed Content Date	Jan 1, 2018
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	electronic

Print This Page

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Mar 04, 2020

This Agreement between Bijan Fakhri ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4782260155066
License date	Mar 04, 2020
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	A Low Resolution Haptic Interface for Interactiv Applications
Licensed Content Author	Bijan Fakhri, Shashank Sharma, Bhavica Soni et
Licensed Content Date	Jan 1, 2019
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	electronic

Print This Page