Representations for the Blind of Depth Information in Photographs

Émmanuel Wilson

Department of Electrical & Computer Engineering
McGill University
Montréal, Québec, Canada

April 2025

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Science.

2025/04/15

Abstract

The Internet, a primarily visual medium, presents challenges for blind and low-vision (BLV) individuals, especially with interpreting commonplace web items such as online graphics. Traditionally when viewing photographs, the observer depends on their understanding of visual concepts and perspective to make sense of a scene, which is inherently problematic for a blind individual. This research aims to bridge this accessibility gap by developing an automatic pipeline that leverages the BLV community's perspective of the world and apply it to photographs using a consumer-affordable device. In addition, this research aims to explore the effects of photograph interpretation for the BLV community using a camera-based view (front-facing) or floor-plan (top-down) perspective.

The main study presented in this thesis uses a 2-DoF force-feedback device, 2DIY from Haply Robotics, to simulate the placement of objects in photographs within a virtual space. Data were collected from seven BLV participants from the Greater Montreal Area. They explored four photographs (two from each perspective) and provided feedback through semi-structured interviews. Findings indicated that most blind users preferred the top-down perspective when exploring photographs with participants giving this perspective a greater confidence score, although not statistically significant. Additionally, it was found that participants could parse foreground and background objects within the scene and infer more details regarding the photograph using the top-down perspective than those using the front-facing view. These results provide design considerations for future studies, as well as the basis of a pipeline, supplying researchers with the tools required to convert a graphic from a front-facing view to a top-down one.[1] The observations made support findings from literature which suggest that the BLV community's interpretation of certain spaces, such as rooms or large spaces, are represented from a top-down perspective. The results from this thesis further suggests that this perspective can also be applied to space within photographs. This opens the door for future research to expand accessibility to online graphics for the BLV community by how one represents spatial information in a photograph using a novel approach.

[1] https://github.com/Shared-Reality-Lab/IMAGE-server

Résumé Scientifique

L'Internet, un média principalement visuel, présente des défis pour les personnes aveugles et malvoyantes (BLV), notamment pour l'interprétation d'éléments courants du web comme les graphiques en ligne. Traditionnellement, lors de la visualisation de photographies, l'observateur dépend de sa compréhension des concepts visuels et de la perspective pour comprendre une scène, ce qui est fondamentalement problématique pour une personne aveugle. Cette recherche vise à combler cet écart d'accessibilité en développant un pipeline automatique qui exploite la perspective du monde de la communauté BLV et l'applique aux photos en utilisant un appareil abordable pour les consommateurs. De plus, cette recherche vise à explorer les effets de l'interprétation des photos pour la communauté BLV en utilisant une vue a base de caméra (vue de face) ou une vue en plan (vue de dessus).

L'étude principale présentée dans cette thèse utilise un dispositif à retour de force à 2 degrés de liberté (DoF), le 2DIY de Haply Robotics, pour simuler le placement des objets dans les photos au sein d'un espace virtuel. Des données ont été collectées auprès de sept participants BLV de la région du Grand Montréal. Ils ont exploré quatre photos (deux de chaque perspective) et ont fourni leurs avis avec des entrevues semi-structurés. Les résultats ont indiqué que la plupart des utilisateurs aveugles préféraient la vue de dessus lorsqu'ils exploraient les photos, les participants attribuant à cette perspective un score de confiance plus élevé, bien que cela ne soit pas statistiquement significatif. De plus, il a été constaté que les participants pouvaient identifier les objets au premier plan et à l'arrière-plan dans la scène et déduire plus de détails concernant la photo en utilisant la vue de dessus par rapport à ceux utilisant la vue de face. Ces résultats fournissent des considérations pour les études futures, ainsi que la base d'un pipeline, fournissant aux chercheurs les outils nécessaires pour convertir un graphique d'une vue de face à une vue de dessus.[2] Les observations faites soutiennent la literature qui suggèrent que l'interprétation par la communauté BLV de certains espaces, tels que les pièces ou large espace, est représentée depuis une perspective vue de dessus. Les résultats de cette thèse suggèrent également que cette perspective peut être appliquée à l'espace au sein des photos. Cela ouvre la voie à des recherches futures visant à étendre l'accessibilité aux graphiques en ligne pour la communauté BLV en explorant de nouvelles approches pour représenter les informations spatiales dans une photo.

[2] https://github.com/Shared-Reality-Lab/IMAGE-server

Acknowledgements

I want to give a special thanks to everyone who supported me throughout my master's. To my parents, siblings, and family for being incredibly loving, patient, and understanding throughout the entire process. I could not have done this without you all. To my friends for being present during the good times and providing me with more support than anyone could ever hope for during the hard times. Thank you to Steve, Jen, and Jacob for providing me with a home away from home, your family kept me grounded and it has been a pleasure to be a part of it. Thank you to Len, whose invaluable support I could not have gone without throughout this journey. To my amazing coworkers in the IMAGE project team and the SRL lab, whose goodwill and positivity cannot be understated. I would like to specifically thank Cyan and Juliette for being incredible colleagues and teaching me so much. I attribute much of my success to our many talks, brainstorms, and support. I would also like to thank Jeff for guiding me through my academic and professional career, answering my many questions and entertaining my discussions. Thank you to Rohan for putting me on the path that ultimately led me to this thesis and thank you to Max for your eloquent reviews and corrections. Finally, thank you to Prof. Jeremy Cooperstock for giving me the opportunity to pursue my master's and allowing me to grow both as an academic and as a person. Your diligence, determination and work ethic is an inspiration to us all.

I acknowledge that this thesis used Grammarly and ChatGPT AI tools for corrections, translations, and sentence smoothing. All the original text was created by the author.

Contents

List of Figures

List of Tables

Acronyms

AI artificial intelligence.

alt-text alternative text.

BLV blind and low vision.

DoF degree of freedom.

FFD force-feedback device.

IDE integrated development environment.

IMAGE Internet Multimodal Access to Graphical Exploration.

JSON JavaScript Object Notation.

ML machine learning.

REB research ethics board.

ROI region of interest.

TTS text-to-speech.

VR virtual reality.

Chapter 1

Introduction

The Internet is a powerful tool accessed by billions of people across the world [1]. It is an important method for accessing and sharing information between people [1]. This is also true for the blind and low vision (BLV) community, who regularly depend on this medium on a daily basis for work and for leisure [2]. As years have passed, the use of online graphics has drastically increased on commonly accessed websites, resulting in decreased accessibility for members of the BLV community [3, 2]. This research aims to work towards a novel solution for approaching online graphics for the BLV population through affordable consumer hardware and the application of perspective, force-feedback, and audio.

Currently, the most common method of interpreting online graphics for the BLV community is through alternative text captions (alt-text) [4]. These captions are populated by the graphic's authors who intend to describe the photo's contents for BLV individuals. However, author participation is not guaranteed and often results in incomplete, uninformative, or missing captions in online graphics [3]. Audio is also used to inform BLV individuals of photograph content. For example, text-to-speech (TTS) is most commonly used to verbalize the alt-text content generated for online graphics [4]. Other methods of sonofying visual content have also been created, such as audiocons which mimic real-life sounds, to earcons which are abstract synthetic audio used to convey specific information [5]. Hardware solutions, such as tactile graphics and swell paper, exist to allow BLV individuals to touch photographs on specialized paper [6, 7]. However, the tools to create these types of tactile mediums are quite slow, and access to them is prohibitively expensive [8]. Refreshable braille displays can also be used to create similar tactile displays at a lower resolution, but they too are overly expensive for a regular consumer [9]. On the other hand, force-feedback devices (FFDs) can generate a wider range of haptic feedback to the user at a higher resolution [9]. FFDs, much like other hardware options available, are also generally expensive. However, the 2DIY from Haply Robotics offers a competitive price point, which opens the door to consumer-available experiences for the BLV community.

Research has been done studying the BLV community to gain a better understanding of their perception of the world [10]. Current literature finds that BLV individuals employ many strategies to perceive the world around them. For large objects and spaces, however, it is suggested that the BLV community opts for a top-down perspective when representing space, creating a floor plan of the intended area [11]. Further research was also conducted studying the BLV individual's perception of photographs [12]. 2D projections such as photographs have long been a pain point for members of the BLV community, especially for the congenitally blind who were born without sight [2]. On their own, graphics are not very informative to a blind individual, but if provided with enough context, the BLV community is able to gain some understanding of its contents [13]. Even with appropriate context provided to the user, photographs are still inherently built with vision in mind. Little research has been done to transform online graphics from a vision-based task to a spatial one, denying members of the BLV community access to photographs that cater to their mental representation of space.

The research outlined in this thesis aims to explore a novel approach to photographs for the BLV community using perspective.

Specifically, our contributions are to:

- Provide novel insight into perspective effects, specifically, how the use of different perspectives impact
- Propose and implement a novel approach towards an automatic pipeline for generating top-down perspective graphics from front view pictures to enhance the spatial understanding of a scene for the blind population.

This thesis provides the basis on which to build a system that allows BLV individuals to gain a greater understanding of the contents of a photograph and where objects are placed within the space. The collected results will better inform designers from which perspective they may wish to present photographs and how to present the information to better benefit the BLV community.

Representing a 3D scene within a 2D workspace presented a challenge when designing experiences for these studies. One such challenge was the application of non-natural mapping of the haptic workspace and its virtual representation. This led to the outcome of movements within the physical workspace not matching that of the virtual outcome in certain experiences; for example, movement along the Z-axis in the physical world might be mapped to movement within the Y-axis in the virtual workspace. Another limitation is the fact that the objects within the virtual scene were limited to a simple circular shape, which normalizes the user experience at the cost of scene fidelity. The haptic representation of objects was also normalized, making all objects feel the same, ensuring each object is equally salient to the touch at the cost of relaying additional information through haptics.

The work outlined within this thesis begins with the literature review, which covers previously conducted research and current tools available for the BLV community to access photographs. This is followed by a chapter describing the design choices made throughout the study. This chapter covers system tools and construction of the IMAGE pipeline, the initial study design, and preliminary results from initial testing, followed by design iterations. The next chapter covers the final version of the study, how it was administered, and discusses the final results. Finally, culminating in the conclusion chapter.

Chapter 2

Background

2.1 Preface

Numerous efforts have been made in both academic and industrial sectors to address the accessibility of online graphics for the BLV community through image captions, sonification, and hardware solutions such as refreshable pin arrays or 3-DoF devices [9]. Despite this, the BLV community's need for additional information on online graphics has yet to be satisfied [9]. This chapter will outline the perception of space and haptics, the interpretation of photographs for blind and low-vision people, and the current tools available to them.

2.2 Perception and Space for BLV community

2.2.1 Space and Navigation

Navigating space on a day to day basis without sight has a profound effect on how someone may perceive the world. A blind individual relies purely on their sense of touch, hearing, and movement to gather information about their surroundings [14]. To take full advantage of the senses available to them, orientation and mobility (O&M) specialists are employed to teach members of the BLV community how to leverage their hearing and sense of touch for navigation [14]. Despite lacking visual information, BLV individuals have a capacity for spatial memory and navigation that is on par with the sighted population [15]. Given that BLV individuals have a capable spatial memory, they would be able to build a spatial map of objects or landmark locations within a virtual scene, such as a photograph. On the other hand, for inference-based spatial tasks such as creating shortcuts between two known locations, it was found that BLV individuals do not score as highly as the sighted population on average [15]. Due to the difficulty in making spatial inferences, BLV individuals may have trouble relating objects or landmarks to one another.

2.2.2 Perspective

Although the mechanisms by which the blind gather information differ from sight, there are many parallels in the information available between vision and the sense of touch. As a result, the ability to see or feel an object allows both sighted and BLV individuals to come to similar conclusions regarding its properties [10]. Through their sense of touch, BLV individuals typically build a three-dimensional mental model of whatever they interact with. When asked to draw certain objects, such as chairs, cubes, tables, etc., blind individuals often depict a top-down, profile or “fold-out” view (where every side of the object is illustrated as if the shape has been flattened and folded out) of the object but rarely consider the effects of perspective, such as the shrinking of a distant shape, in their interpretations [16, 17]. This lack of experience in the effects of perspective makes photograph interpretation more challenging, as a BLV individual would have difficulty differentiating between a small object or one that is at a distance [18]. This is particularly relevant when approaching photographs from a front-facing view, where these sorts of perspective distortions, which stem from a vision-based approach, are most common. While depicting certain large objects and spaces such as a bathtub or a room, blind individuals appear to favour a combination of a bird’s eye (or top-down) and fold-out view, where occluded features such as the legs of a table are splayed out to indicate important information while simultaneously creating a floor plan style representation of the space that better aligns with navigation strategies [11]. The importance of this type of representation of space is further supported by the observation that blind children typically orient themselves through landmarks as opposed to their sighted counterparts, which orient themselves based on their current position in space [19]. Altogether, this would suggest that a top-down representation may be more informative to the BLV community for spatial-based tasks, while a profile or front-facing representation would be more informative for certain details of objects.

2.2.3 Context

Due to the common angle in which most photographs are taken, graphics are typically presented in a front-facing view. Despite this being a common perspective, if a blind user is provided with a tactile graphic of a scene or object in the front-facing orientation, participants were found to have difficulty identifying objects without context [13]. However, if a context was provided prior to scene exploration, BLV users were found to be very reliable at identifying the graphic's contents [13]. This shows that context is essential for interpreting tactile graphics and allows users to fill in gaps in their understanding of a scene. In order to push participants to rely on the position of objects within the haptic space in a given orientation, this work aims to provide the minimum context required to construct a mental representation of a scene.

2.2.4 Perception of Haptic

The loss of vision leads BLV individuals to have to rely on senses other than vision in order to navigate daily life. As a result, it has been found that members of this community have developed a keen sense of touch and are able to process this information faster than the sighted population [20]. Designers and researchers take advantage of this and create ways to convey tactile information to users through haptics [9]. This can come in the form of heat, proprioception, texture, vibrations, and forces [21]. Most commonly, vibrations, textures, and forces are used for accessibility technologies to allow BLV individuals to improve their quality of life [9]. These haptic features are also employed throughout the work presented in this thesis to convey spatial information about objects within photographs.

2.3 Accessibility Approaches for Photographs

Accessibility of photographs for the BLV community is hardly a new topic within the HCI community. Throughout the years, there have been numerous developments in accessible technology for the BLV community [7]. Despite this, access to online graphical content for BLV individuals is far from a solved problem [7] With the increasing prelevance of graphical online content [3], further development of alternatives for visual information for the BLV community is essential [4].

Currently, multiple strategies are available for people in the BLV community to approach photographs they encounter online. These solutions can be split into two categories: software, which includes alternative-text (alt-text) and sonifications, and hardware, which includes braille displays, tactile graphics, and force-feedback devices(FFD).

2.3.1 Software Solutions

2.3.1.1 Alt-Text

Alt-text has long been the most common method of making online graphics accessible to the BLV community whereby the content author provides a written description of the graphic in question. The efficacy of this method, however, is strongly tied to the diligence of the author to provide these notations. Sadly, about a third of the images found on popular home pages have missing, questionable, or repetitive alternative-text [3]. The presence of online graphics has seen an increase of 28% in the last year, meaning that the gap left by missing or inadequate captions will have an even greater effect on the BLV community [3]. To make up for the lack of author participation, machine learning algorithms (ML) are often introduced to automatically generate alternative-text for a more complete description of online graphics [22] Our work also uses ML tools to extract the relevant objects within the scene, identify them, and determine where they are located within the space. It is a promising solution, however with some limitations. Correct scene interpretation from the ML algorithms, as well as the creation of human-readable and accurate scene descriptions, are non-trivial tasks, as described in MacLeod et al. [23]. Small mistakes in the automatically generated text can often mislead blind users without a way to validate the scene for themselves [23]. We believe that a BLV individual may be able to validate the description of a photograph by pairing ML-generated object tags with a haptic spatial representation of a scene.

2.3.1.2 Sonification

Another commonly used medium that is used to convey information to a user is audio. This method of communication is often tied to alt-text where the generated text is read out to the participants. Since this kind of audio output requires no prior knowledge, its use is explored in this work along with other types of audio. There are also other non-speech methods of presenting information through sound which is called “Sonification”.

It is possible to encode photographs and paintings with sound as shown in Anderson et al. [24] and Jorge et al. [25], where they transformed the visual medium into music. It was found that human users struggle to accurately decode precise information regarding the pictures/paintings but have a better grasp of the emotion conveyed by the music. An alternative method of leveraging the layering capabilities of musical instruments is to map the colour of pixels to instruments, as shown in Banf et al. [26]. In this study, colour and saturation correspond to the sound of a particular instrument, and a transition between colours will result in a transition between instrument outputs. Using this method, participants were able to accurately determine pixel colour and the transition of colour throughout a graphic; however, they had difficulty determining the graphic’s actual contents. Rather than encoding the entire colour spectrum as they did in Banf et al. [26], our work explores the benefits of mapping physical properties of the scene, such as depth, to sound in order to convey the position of objects.

To help reduce the abstraction of sonification, earcons and audio-icons can also be used as audio cues within a soundscape, which can provide additional information. Audio-icons are audio approximations of naturally occurring sounds in everyday life, such as a cow mooing to indicate that there is a cow within the photograph [5]. They are often used for non-speech audio analogies and are most commonly found in children’s toys, but they have also been used in research for the BLV community. Audio-icons usually depend on the user understanding the analogy they represent, and some audio-icons may sound quite similar to one another which may lead to misunderstanding or confusion [5]. Earcons are similar to audio-icons but more abstract and are synthetically generated and manipulated to present information. This form of non-speech audio requires no previous world knowledge from the user and can be shaped by the designer to present specific information with optimal contrast. This does, however, require context as to what the sound represents to make sense of what is being presented. Although there is no official standard for what properties non-speech audio can represent, there are often physical associations that should be taken into account [27].

- Loudness is associated with distance
- Pitch is associated with vertical position
- spatialization associated with horizontal position

These physical associations to sound, although common, are not always used to map those specific properties in research or industry, as certain tasks may require different audio features [27]. Earcons will be used in this study as they have the ability to be quick and concise, which lends itself to the rapid presentation of information in small doses. This will allow BLV users to explore a photograph and rapidly collect information regarding an object's position within the space.

Sonification is a very powerful tool available to create immersive and impactful experiences for the BLV community. It is typically inexpensive to implement and easily available, only requiring speakers or headphones. Audio, however, is a hotly contested medium for a BLV user. Since visual information is not an option, it often falls to audio representations to pick up the slack, as hardware solutions are not always available. Because of this, the soundscape is often in competition with itself, and designers must be careful not to overwhelm users by presenting too much information. With the majority of the BLV population being elderly (65+ years) [28], certain sonification strategies might not be effective, as hearing loss is also a common condition for people of advanced age.

2.3.2 Hardware Solutions

2.3.2.1 Braille Displays and Tactile Graphics

Alt-text and sonification strategies offer many advantages, but they are limited to a single channel for information transfer. Hardware solutions, such as tactile graphics and braille displays, provide information by leveraging the sense of touch and proprioception.

Swell paper is an easy-access tactile medium that allows people to draw on it with a specialized marker whose lines then rise when exposed to heat to create a custom tactile graphic. This allows researchers and others to create personalized graphics for members of the BLV community or even allow people with low vision to create their own drawings. A systematic way to generate graphics using swell paper was created using object detection algorithms to create a collage of icons to represent the contents of a picture [6]. This was later developed into Pic2Tac [29] that can extract information from a picture and create a collage of symbolic textures and icons on swell paper. This includes background information and object icon representation, which allows a BLV individual to feel a simplification of the graphic. The objective of this thesis is to work towards a comparable goal by utilizing forces and audio instead of relying solely on a tactile medium.

Similarly to swell paper, embossed paper is a physical medium that is marked in order to create raised lines. The biggest difference between the two is that embossed paper is physically pushed out to make lines and shapes rather than swollen to do so. Historically, this is the most common form of tactile displays available for the BLV community, however access to the facilities to create detailed embossed graphics is prohibitively expensive [8].

Tactile graphics have been shown to be quite an effective method of presenting photographs to BLV individuals so long as they are presented with context describing the scene [13]. Without context, the shapes and outlines are difficult to interpret on their own for a blind user. A downside in using a physical medium, such as swell paper and tactile graphics, is the time it takes to produce them. Not only that, users are also required to purchase the tools needed to produce these graphics, as well as the added cost of resource replenishment and subsequent storage of the used sheets. In contrast, purely digital interfaces, such as Braille displays, touchscreens, and FFDs generally incur only an initial acquisition cost and typically do not generate ongoing expenses during their operational lifespan.

Using a rising and falling array of pins, the refreshable braille display is able to rapidly generate braille text but also create tactile displays. As seen in Zeng et al. [30], refreshable braille displays can be used to depict graphics such as maps. A major advantage of tactile graphics and refreshable braille displays is the ability to touch multiple points along the graphic with multiple fingers or hands. This affordance to touch multiple places along the workspace at once allows BLV users to quickly scan the space and create a strong spatial relationship between two points [31]. However, a major disadvantage of braille displays over tactile graphics is the limited resolution of the pins within the pin array. Due to this limitation, refreshable braille displays often resort to panning and zooming to overcome the lack of resolution [30]. These kinds of devices are also prohibitively expensive and, therefore, are limited to institutions such as libraries and schools as opposed to the individual consumer. These devices were not considered for this experience because of the limitation of the haptic rendering quality and the generally inaccessible price point. Not only are these devices expensive, but braille literacy has also been on a steady decline, with only about 12 % of the blind population being braille literate [32]. This only reduces the usefulness and potential benefits provided by this type of device, as they are less versatile compared to touchscreens or FFDs that require little or no prior training [33, 34].

2.3.2.2 Force-Feedback

All previous described strategies and devices lack the application of force, which is possible when using a force-feedback device (FFD). FFDs come in many forms and can offer different degrees of freedom (DoF) depending on the user's needs. These devices often use the form factor of an arm(s) attached to a motor(s), allowing the user to move the limbs along the device's available axes and are used to apply force to the user's hand at the end-effector. This form of haptic representation affords a high-resolution experience along a single point of contact, able to convey a sense of texture, dynamic physics, and active guidance of the user. Force-feedback devices, however, appear to be one of the least widely used tools within the community, seeing a dramatic loss of interest in the decades since the release of the "Phantom Omni" [9].

Devices with a single degree of freedom (1-DoF) are quite limited in what they can represent, as they only offer mobility and forces along a single axis. Despite this, they have seen use in virtual reality (VR) applications to provide an immersive experience to their users, bringing some limited virtual experiences into the real world [35]. However, these kinds of devices are too restrictive for the purposes of graphical representation that we wish to convey.

3-DoF devices, on the other hand, such as the Phantom Omni, are able to articulate and exert force along 3 axes and simulate physics in 3-dimensions. This opens the avenue to many possibilities and has been used in the field for representing graphics and virtual spaces. 3-DoF FFDs have been shown to provide convincing texture renderings and can be used to identify simple virtual shapes [36]. However, with a single point of contact, more complex shapes or environments become increasingly more challenging and time-consuming to identify. The Phantom was also used to explore photographs as seen in Lareau and Lang et al. [37] where they separated the photograph into a hierarchy of contour segments and 3D texture for the user to explore that made photographs more understandable for its blind participants [37]. Despite the fact that the 3-DoF is such a powerful tool, the price point of such a device outweighs the potential benefit offered to an individual member of the BLV community.

Unlike a typical 3-DoF device, a 2-DoF device, such as the 2D pantograph from Haply Robotics called 2DIY, has a drastically lower price point that is within reach for the regular consumer. Although limited to two axes, 2-DoF devices are capable of generating complex force-feedback to their user and offer the minimum dimensionality required to represent a photograph. Blind participants tasked with identifying cancerous cells in histological graphics were found to be more accurate when using an FFD over a tactile graphic, although it took longer [38]. This shows that despite the limitation of a single point of contact, the richness in texture provided by FFDs can prove to be highly beneficial. A similar paradigm was used using ultrasonic vibrations while exploring a touch screen as seen in Tividar et al. [39], where a person explores the layout of a room and can identify aspects of the room through vibration. Although similar, the FFDs used within our works will allow for the application of force and vibration rather than only vibration. This not only opens the door to a greater variety of haptic effects but also allows a participant to be guided within the space, such as pulling their hands towards a target or pushing them away from another.

Chapter 3

Haptic Representation of Photographs

Preface

Members of the BLV community live their lives in a three dimensional world, and as such have difficulty understanding the two dimensionality of photographs. Presenting photographs using a 2D pantograph, such as the 2DIY, requires various techniques that were explored in order to generate a 3D scene using a 2D device. In this chapter, we are looking to expand our understanding of a BLV individual's representation of a virtual 3D space generated from a single image. This section outlines the design considerations and challenges of creating a study for the BLV community for a 3D representation of space using a 2-DoF device.

Author's Contribution

Study design of 3D representation of space from a photograph using a 2-DoF device. Co-authors, Cyan Kuo and Juliette Regimbal were involved in the ethics, study design and testing process.

3.1 Introduction

Graphical content, such as photographs, have become an integral part of the online experience in recent years. Despite the increasing presence of photographs for day-to-day interactions on the Internet, accessibility for this type of content is still a pain point for the BLV community [4]. In its current state, the online graphics experience for a blind individual has many shortcomings, as described in the previous chapter. 2D projections of a 3D scene, such as a photograph, are particularly challenging for a blind individual to interpret. This is especially true for someone who is blind from birth as they lack any previous experience with perspective and have only experienced a physical 3D world through touch and sound.

Through the use of machine learning, the three-dimensional information of a photograph can be untangled from its 2D form. Using this additional information, a new 3D representation of a photograph can be generated and presented through audio and haptics. Since 3-DoF devices are prohibitively expensive, an approximation of a 3D scene is made using a 2-DoF device. Despite losing some fidelity in the haptic scene when using a 2-DoF device in such a way, doing so would greatly lower the barrier of access for anyone within the BLV community due to its lower cost. This chapter includes the design choices and study iterations made to uncover the BLV community's preferences in perspectives when presented with photographs using a 2-DoF device.

3.1.1 Apparatus

Over the course of this experiment, the 2DIY from Haply Robotics, as seen in Fig. 3.1, was used to deliver forces to the user as they interacted with the virtual graphic. The 2DIY was a 2-DoF pantograph device containing two motors. The motors actuate each arm individually, applying forces at the point where they meet, the end-effector. Through forward kinematics, the device administers the appropriate torque in the motor to accurately apply forces to the end-effector. Reverse kinematics was also performed by measuring the angle of arm rotation within the motors to calculate the physical position of the end-effector. Through these calculations, the device could track the user's movements and apply forces when needed to create an immersive haptic experience.

With its relatively simple construction and components, the 2DIY was quite affordable to

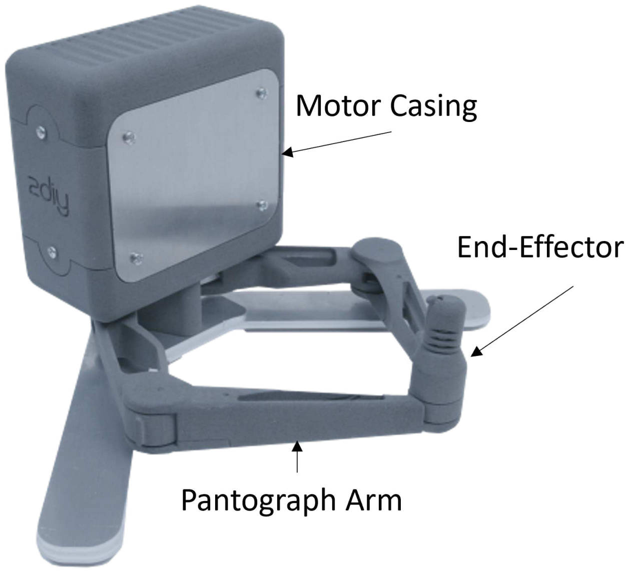


Fig. 3.1: 2DIY Gen3.1 used during this experiment.

the general public, being sold for approximately $500 prior to this experiment. In comparison, similar devices and other accessibility technologies were typically priced in the thousands of dollars. However, the 2DIY was not without its limitations. Notably, it was limited to two degrees of freedom, whereas most similar FFDs opt for three degrees of freedom, opening them up to more use cases compared to the 2DIY. With its small size, the user could also easily overpower the motors. Although it could provide a wide variety of forces, there is a limit to how much power they are able to exert.

All of the experiences were generated in a Java-like integrated development environment (IDE) called Processing. The object positions and names were generated through the IMAGE web extension.

3.2 Systems and Tools to Construct Representations

The first step in creating a virtual experience for photographs is to extract the relevant information needed from online graphics. The IMAGE web extension created by the SRL Lab at McGill

University is a free to use Chrome extension that is designed to provide online access to graphics for the BLV community [40]. This Web extension is able to extract an online graphic chosen by the user and process the graphic through a hierarchy of machine learning algorithms (called preprocessors), whose outputs are then collected and presented to the user through handlers. Handlers curate the preprocessor outputs and render the results through audio. Currently, this includes spatialized audio of objects (panning left to right), spatialized audio of the contour of an element of the scene (such as the sky), as well as a verbal presentation of the scene's contents. All of these results are automatically generated and take only a few seconds from graphic extraction to user presentation.

The IMAGE extension architecture is designed with modularity in mind and affords designers the possibility to add additional functionality to the web extension. In order to create a 3D environment out of a 2D photograph, additional preprocessors and handlers are required. Traditionally, depth extraction of photographs has been done using multiple cameras/angles for stereo matching [41] or sensor based approaches such as using depth cameras. However, given the information available from online content, these depth extraction techniques cannot be used on single frame photographs. With the use of machine learning, depth extraction of a single still frame graphic has been made possible. There are a number of available monocular depth extraction algorithms to choose from, but they must meet certain criteria.

1. Perform purely monocular depth estimation (i.e., Requires no additional components)
2. Provides pretrained weights
3. Licensing that allows commercial use
4. Fast processing time (less than three seconds)
5. Low resource consumption (16 GB RAM and 6 GB GPU)

In order to be able to process any photograph available online, the algorithm must be able to extract depth using only the photograph, as additional information required by other algorithms will be nonexistent. The machine learning algorithm used should also have pretrained weights available to maximize performance and avoid potential bias or overtraining. Since the IMAGE project is open-source and freely available, the algorithms require a license that allows other researchers and commercial partners to leverage this addition to the IMAGE extension. At the time of writing, there were four potential candidates that satisfied the first three criteria. These were Adelai Depth [42], GCN Depth [43] as well as Boosted Midas and Boosted LeRes depth algorithms [44]. These algorithms were evaluated on a small battery of graphics to measure their performance, processing time, and resource consumption. All tests were performed on an AMD Ryzen 7 with 32 GB of DDR4 RAM and Titan Xp 12 GB GPU.

The resulting processing time and resource consumption of the tests, as seen in Table 3.1, show that Adelai Depth is the most efficient algorithm out of the potential candidates. Since this is meant for a real-time application, the response time of the IMAGE extension is mandated to be within three seconds to avoid long waiting times for an end user. Out of the four models, only GCN Depth and Adelai Depth fit within these parameters, with Adelai Depth taking a fraction of the time to process. The Boosted algorithms (both Midas and LeRes) take far too long for real time application. Processing time seems to correlate as well with the resource consumption of the algorithms. Not only is Adelai Depth the fastest algorithm by far, but it also consumes the least amount of RAM and GPU. As the IMAGE extension has finite resources available and runs multiple machine learning preprocessors simultaneously, mitigating the computational stress applied to the system by the addition of this preprocessor is critical. For this reason, the depth estimation preprocessor should consume within 16 GB of RAM and 6 GB of GPU, to provide adequate resources for the rest of the system.

Table 3.1: Resource comparison across depth generating ML models.

Depth Estimator	Processing Time (sec)	RAM (GB)	GPU (GB)
Adelai Depth	**0.3**	**3.7**	**0.7**
GCN Depth	3	4.5	2.5
Boosted Midas	62	5.8	10.3
Boosted LeRes	47	5.2	7.3

The resulting outputs of the tests, seen in Fig. 3.2 show the quality of depth extraction for each algorithm. Despite the high processing time and resource consumption, Boosted Midas appears to struggle at extracting depth from further within the photograph space, such as the background hockey players in the third photo on the right. GCN Depth also appears to have a lot of artifacts with its resulting depth extraction. The front pillars and roof of the gazebo appear to be at very different depths when they should be more similar, the depth of the car in the last picture is also poorly extracted (i.e., the actual shape of the back of the car does not match the shape of the extracted depth), and the hockey example is completely unintelligible. Adelai Depth has a reasonable depth extraction quality. Certain details, such as the very back shelves of the first picture and the very front bushes of the gazebo picture, are not clearly picked out. However, the overall depth extraction encompasses all of the most salient features, such as the car in the last picture and even the background hockey players in the third picture(from left to right). Finally, the highest quality of depth extraction seems to be generated by Boosted LeRes. Although slow, this machine learning algorithm is capable of generating a detailed depth representation of the graphics. Considering the processing time, resource consumption, and quality of depth extraction, the Adelai depth machine learning algorithm is the clear choice for our performance targets.

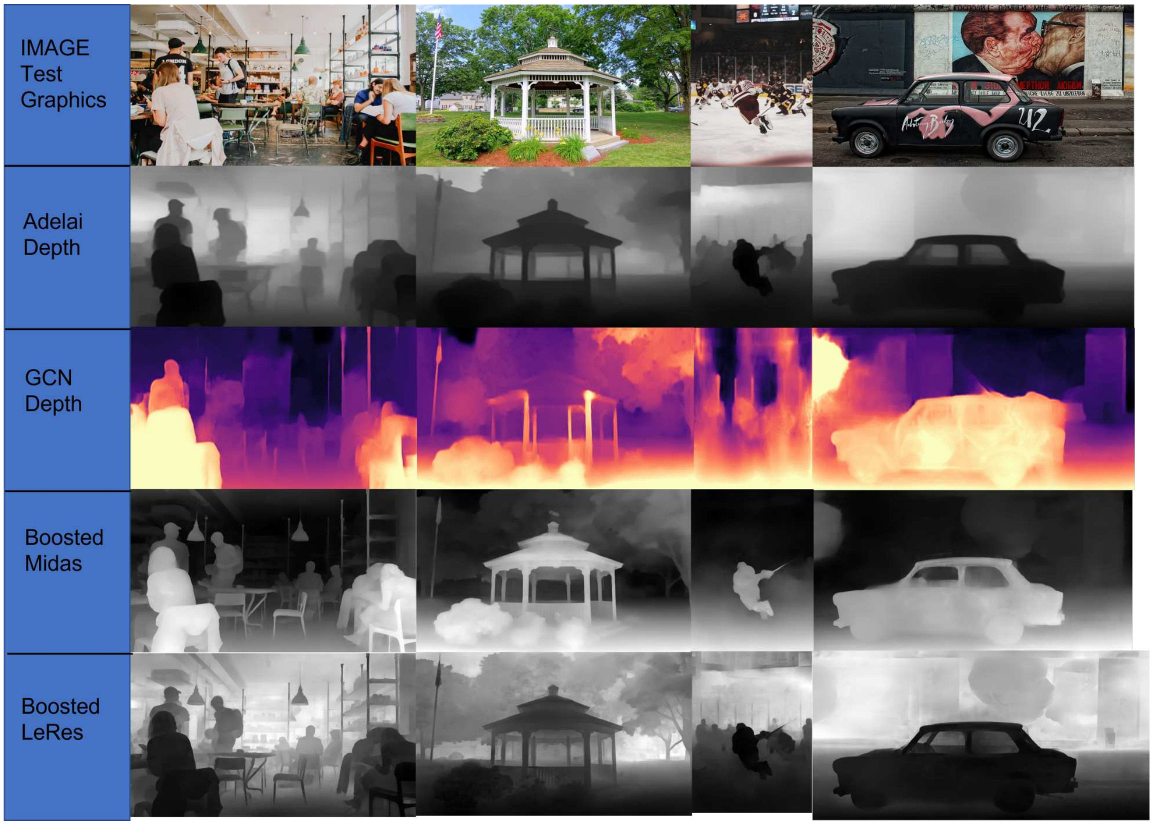


Fig. 3.2: Depth generation quality comparison across algorithms.

After integrating the depth preprocessor into the web extension, the ML algorithm would save the photograph output as a normalized grayscale image with depth values from zero to one into the IMAGE request JavaScript Object Notation (JSON) package. This JSON is then passed through to the rest of the IMAGE architecture, allowing the preprocessed output to be used by other preprocessors and handlers. In order to extract useful depth information about the scene, an additional preprocessor was created. Largely inspired by Masoumian et al. [45], the depth extraction information was paired with the YOLOv8 object detection machine learning algorithm to extract the approximate depth of objects within the scene. Only objects detected using YOLOv8 with a confidence of 75% or greater were considered. The region of interest (ROI) of each object generated by YOLOv8 were then overlaid on the depth image. The median depth pixel value within each individual object's ROI was calculated and provides an approximate depth value for each object. Since the ROI does not perfectly encapsulate the detected objects, the median depth pixel value is chosen to avoid the surrounding background values that may pollute the depth approximation. A debug handler was also created to visualize the depth output as a sanity check for the quality of the depth extraction.

In future work, a handler will be created that will automatically populate the haptic space within a 2-DoF device with the scene objects and background information for a user to explore. For the purpose of this experiment, this handler is approximated by manually creating the haptic experiences through Processing scripts. Each object included in these scenes shares the same three forces applied by the 2DIY when entering their shapes. Firstly, a spring force is applied on each object that serves to help find them during a quick pass of the end-effector during initial scene exploration. Secondly, each object also receives a texture and friction force to help indicate to the user that the end-effector is still located within the object's ROI during exploration.

3.3 Study Design

3.3.1 Initial Study

This section outlines the design choices of our initial quantitative experiment. The goal is to measure and evaluate the effectiveness of representing 3D information across three different conditions. As the 2DIY only has two physical dimensions to work with, the third dimension must be represented through other means. Each condition represents the virtual space in either the front-facing or top-down perspective with audio earcons to indicate the third dimension. Some of these conditions also orient the 2DIY in different positions to explore the effects of natural mapping. Natural mapping occurs when the movements of the end-effector within the physical space match the movements of the cursor within the virtual space. For example, movement along the Z-axis in the physical world would lead to movement within the Z-axis in the virtual workspace. The first condition, as described in Fig. 3.3, has the 2DIY resting parallel to the table. The device's physical workspace represents the X-Z axes and is mapped naturally within the virtual image space, while the Y-axis is represented through audio.

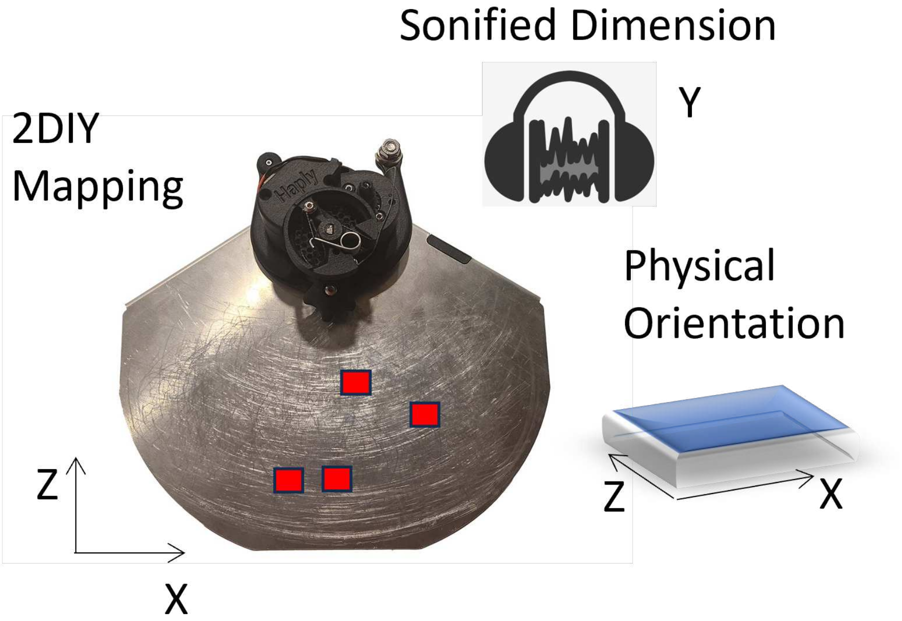


Fig. 3.3: Naturally mapped top-down orientation of the 2DIY. The device is placed parallel to the table surface with its workspace representing the X-Z plane of a photograph while the Y-axis is sonified.

The second condition, as described in Fig. 3.4, has the 2DIY resting in the same position as the previous condition (along the X-Z plane), only the physical mapping of the device explores the X-Y axes of the virtual space while the Z-axis is sonified. This perspective and 2DIY orientation lead to a non-natural mapping of the workspace. Since the movements of the end-effector along the physical Z-axis correspond to the Y-axis within the virtual space. There was concern that this perspective may confuse participants, as making this mental map may incur a greater cognitive load.

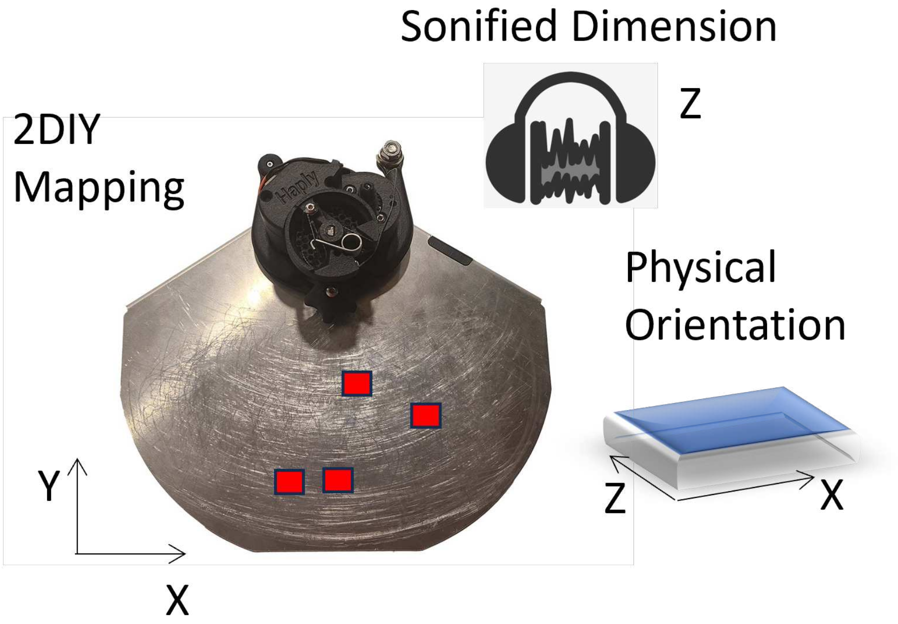


Fig. 3.4: Front-facing orientation of the 2DIY. The device is placed parallel to the table surface with its workspace representing the X-Y plane within the photograph while the Z dimension being sonified.

The third and final condition, as seen in Fig. 3.5, has the 2DIY hanging perpendicularly to the table such that the workspace is now vertical.

Physical exploration using the device represents the X-Y axes within the virtual image space, and the audio represents the Z axis. In this way, the physical and virtual exploration for the first and third conditions map naturally.

During the experiment, participants explore a virtual haptic space and replicate what they have discovered using LEGO®, as seen in Fig. 3.6.

During the course of the study, each perspective is introduced with a training trial, where a prebuilt LEGO® scene is created and shown simultaneously along with its haptic representation using the 2DIY. The participant is tasked to explore the physical representation along with the virtual one and only pass the training if they can correctly match at least two of the

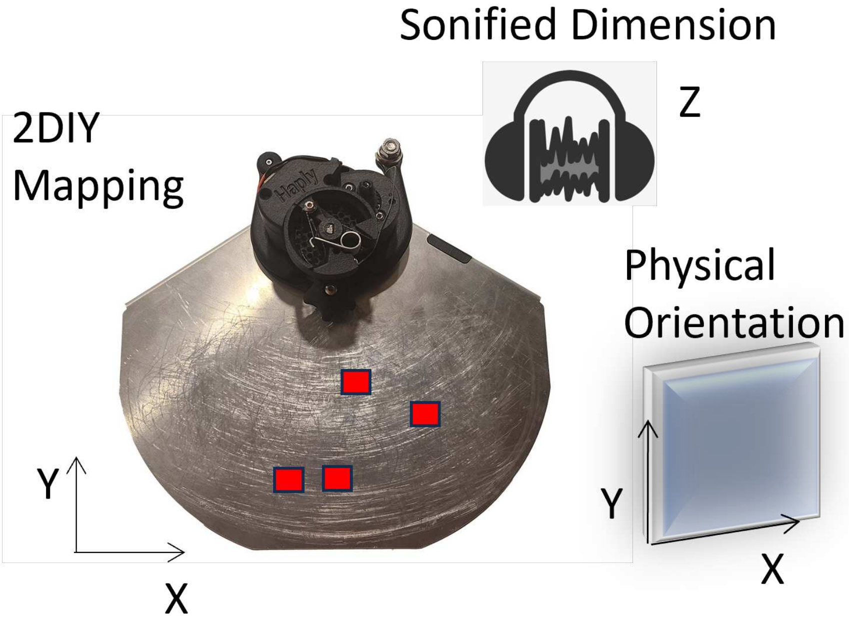


Fig. 3.5: Naturally mapped front-facing orientation of the 2DIY. The device is placed perpendicular to the table surface with its workspace representing the X-Y plane within a photograph while the Z dimension being sonified.

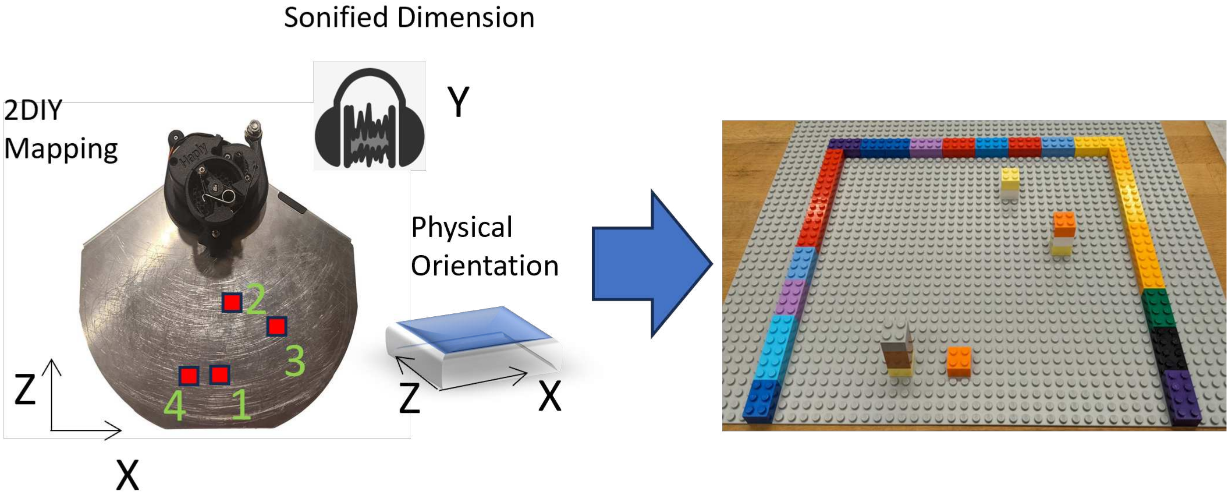


Fig. 3.6: Experiment design: Participant is tasked to reconstruct a scene based on what they feel and hear.

virtual haptic objects with their physical LEGO® counterpart. After completing the training, the participants are presented with a virtual scene on the 2DIY and are evaluated based on how well they can reconstruct the scene using LEGO®. The placement of the LEGO® on the board represents the X-Z axes of the scene, while the height of the LEGO® towers represents the Y-axis. During the reconstruction phase, participants have access to the 2DIY for reference, which we call "parallel exploration". These tests are carried out three times. The fourth and final trial is a memory task, where the user reconstructs a scene from memory and does not have access to the 2DIY during the reconstruction phase. This step is called "sequential exploration" Each perspective is evaluated in the same manner, one training trial, three parallel exploration trials, and a final sequential exploration trial. The scenes presented are balanced across the different perspectives, and each perspective is presented in a balanced order as well. Each condition is evaluated on the basis of whether or not the objects are placed in the correct order along each axis. Relative accuracy along the axes determines which of the three perspectives translates best for the user's understanding of the scene.

3.3.2 3D Representation

As described in Section 3.2, the pipeline built upon the IMAGE extension architecture processes a photograph, extracts the depth information, identifies object coordinates within the pixel space and combines this information to gain an approximation of object location in 3D space. Although just an approximation, this 3D approach to photographs opens an avenue to a new rapid, automatic pipeline for online content. As suggested in Eriksson et al. [18], members of the BLV community interact and interpret a 3D world, and therefore, 2D projections are difficult to decipher. We believed that a 3D representation of the photospace would be most beneficial to the blind and low vision community.

Without adding additional hardware or increasing the complexity of the haptic renderings, the easiest method to represent a third dimension is through audio. Audio sonification has long been used in the field to represent and indicate various properties for the BLV community [27]. As mentioned in Section 2, the audio representation can vary from abstract sounds of earcons to audible text such as TTS. An important factor that had a major impact on the design choices for this project is the fact that a large portion of our target population is elderly. Not only is there a strong correlation with age-associated blindness, there is also a strong association with deafness [46] that should not be ignored. As the general population is steadily increasing in age [28], catering to their needs will maximize the impact of our work.

Sonifications used in our experiment represent the missing dimension, creating an approximation of the 3D space. This audio represents the location of objects along the photograph's Y-axis (vertical position) or Z-axis (depth). Earcons are used and change to indicate a low to high vertical position or near-to-far distance. The following sonification strategies were considered:

- loudness
- temporal difference
- spatialization
- pitch
- timber

Due to its physical association with distance, loudness was considered the third-dimensional representation (depth) for the front-facing view. However, to reduce experimental variance, instead of having two sonification strategies to match its corresponding physical association, only a single audio mapping was chosen and applied across all perspectives. This way, rather than having the possibility of one sonification outperforming another, each perspective would be on equal terms with the same audio cues. With that in mind, due to the fact that a significant portion of the BLV population is also elderly, loudness being used as a third dimension representation was ruled out. If a user is deaf or hard of hearing, determining the difference in volume between two abstract cues is challenging. As they often rely on an increase in volume to compensate for their hearing loss, creating a system that relies on variable volume is unreliable.

Spatialization was also ruled out as an option to represent the height and depth of an object, as it depends on the ability to hear from both ears. This reduces the robustness of this method in a population prone to deafness since any hearing loss is likely to reduce its efficacy.

Another sonification strategy that was ruled out was the use of timber. Timber is a more abstract form of audio, and interpreting the intended mapping of particular timbral characteristics is a non-trivial task that would prove challenging for a naive participant.

Temporal difference, i.e., an object rings for longer if the height or depth value is greater, was also not used. Through empirical testing, this method became less viable the more objects were present in the scene. More specifically, it was difficult to determine the depth or height of any object if the user came into contact with multiple objects within a short time period. This made comparisons between objects time-consuming and difficult, as the audio from one object might still be playing by the time the second object was touched, causing both audios to blend together. This would force a user to wait for the audio to complete for one object before proceeding to the next, which adds unnecessary overhead.

The sonification strategy chosen for the third spatial dimension was pitch. Although pitch is often associated with vertical space [47], it was chosen to represent the third dimension for both the front-facing perspective (depth sonified) and the top-down perspective (verticality sonified). The chosen frequency range was between 1 KHz and 1.5 Khz, specifically chosen to be within the auditory range of most of the elderly population [48]. These earcons were generated using Supercollider for one second sound bites. In order to facilitate the comparison between objects, the frequency steps when creating the pitch audio were generated logarithmically to account for the Weber-Fechner law [49]. This way, higher frequency objects are easier to distinguish from each other as the threshold to perceive these differences increases with frequency [48].

To evaluate this method of 3D representation, standardized haptic environments were created to facilitate comparison between scenes. All generated haptic scenes contain four objects of matching size. Their dimensions were made to match the same ratio as the objects used within the physical reconstruction medium, as seen in Fig. 3.6. The virtual position of each object is required to be at least 10% away from every other object along any axis (i.e., if the length of the x-axis is 10 cm, the minimum distance between objects along the x-axis is 1 cm). This is to ensure that no two objects overlap in space, avoiding potentially confusing or conflicting information regarding object positions.

3.3.3 Preliminary Results and Challenges

A pilot study was conducted on three blind individuals. Unfortunately, only one out of the three participants was able to complete the experiment, while the other two opted out due to time constraints and frustration. The greatest pain points observed throughout these trials were the result of heavy cognitive loads, significant learning effects, and object discoverability. Cognitive load was a challenge due to the large number of scene components committed to memory needed for scene reconstruction. Mental fatigue also played a large factor as the study contained many trials, made worse by the fact that it could take up to three hours to conduct. The transition from one perspective to the next is another important contributor to participants' mental load and frustration. It was clear that a learning effect was taking place, which added to the participant's confusion and overall frustration with the task. So much effort is put into learning the first perspective presented in the study that participants would often replace or hybridize all future perspectives with the first. For example, if the first perspective presented has the X-Y axes along the physical workspace and the Z axis is sonified as presented in Fig. 3.4, then it is likely that when presented with the top-down view from Fig. 3.3 that they would still interpret the scene as front-facing from Fig. 3.4. This misunderstanding would happen even after training, additional clarifications, and participant confirmation. As a result, this mental mismatch of various mappings would only add to the participant's mental load, confusion, and frustration. Finally, another contributor to participant's frustrations is the discoverability of objects. This is due in part to the exploration strategy used by the participants. They mostly followed along the edges of the virtual space, mimicking their real-life spatial mapping strategies taught to them by O&M specialists [14]. Participants would hug the borders of the scene and ground themselves in a corner, using it as a landmark, then venture out from these landmarks in an attempt to find an object. This would often lead to participants taking longer to find objects, as they would venture from the corners to find objects rather than systematically combing the haptic space. They often had a hard time finding the final object within the scenes, which led to further frustration.

Another observation made during the course of this pilot study is the fact that although the physical workspace was placed to map naturally along the vertical axis, as seen in Fig. 3.5, the ergonomics of this method leads to complications. Most of the participants commented that exploring the physical workspace in this manner was awkward. This may be due in part to the fact that holding one's arm in a vertical position while exploring is rather uncomfortable for extended periods of time and also due to the fact that, due to gravity, the 2DIY end-effector moves downward whenever the participant lets go. This, in turn, made the training phase and parallel exploration trials more challenging as most participants used both hands to explore the LEGO® representation. Whenever this occurs, the participants must re-situate themselves and rediscover the location of objects they had previously found.

The LEGO® medium used for scene reconstruction is also not ideal. LEGO® was chosen for its tactile feel, allowing members of the BLV community to feel the spatial units along which they are building along the three axes, X, Y, and Z. This way, participants are able to feel the difference between objects with ease. A problem that was noticed with the LEGO® is the fact that pieces stay in position when pressed together, only when properly aligned. This would cause participants to rotate and shuffle the building blocks into position so that they would fit, often causing them to drift from their intended location. The granularity of LEGO® pegs also creates a forced sparsity for where each piece can be placed, producing a discretized workspace. This, in combination with the drift caused when placing LEGO® blocks, increased the variance in the final position of the reconstructed scene. Similarly, stacking LEGO® blocks is hardly an accurate measure of vertical placement and cannot be compared to the X-Z axes of the LEGO® scene, which have a far greater resolution.

Many lessons were learned from this pilot study. For example, participants got noticeably fatigued by the two hours mark and task frustration, such as unclear training, must be minimized. Learning effects also played a critical role in the task performance and must be mitigated for this study to yield useful information.

3.3.3.1 Second Iteration

With the lessons learned from the pilot study in the previous section, many changes were made to address the underlying issues. Firstly, the experiment was greatly simplified. The third condition, as described in Fig. 3.5, was completely eliminated. This reduced the number of trials while simultaneously forcing a comparison between two perspectives, top-down and front-facing, rather than three perspectives. In order to mitigate learning effects, this new version of the study required two sessions on different days, separated by at least a week. This gap between sessions should drastically reduce the learning effects across perspectives [50], where each day presents only a single perspective.

Further adjustments were made with respect to the training of each perspective, as some of the confusion outlined in the pilot study could have been attributed to unclear instructions. A realization was made that a purely verbal explanation of a given perspective can only go so far as to help the participant's understanding of the task. In addition to clear and concise instructions, a physical aid was provided for the participant to feel the perspective during training, which greatly benefits their comprehension. Unlike the previous study iteration, that expected participants to explore the example LEGO® scene after a brief explanation of the perspective, the training for this iteration focused on breaking down the physical properties represented by the 2DIY using a physical model. As such, example scenes were created to scale with the 2DIY workspace using cardboard, wood, and magnets, as seen in Fig. 3.7 in order to create a comprehensive and in-depth training regime for participants. Magnets were inserted into the bottoms of the wooden dowels and the cardboard, allowing each piece to be fixed and removed from the board at specific locations. This was used in favour of the LEGO® as the magnetic training scenes were able to be built to scale with the 2DIY workspace and were also easier to use as the experimenter, as the magnetic objects could be placed at a glance rather than the meticulous placement needed when building LEGO®. With these modular physical model scenes, the researcher could present the example scenes piece by piece allowing the participant to break down the location of each object individually during training. Gradually, more objects were introduced to increase the difficulty of the scene so that the user could build an understanding of the virtual environment prior to testing.

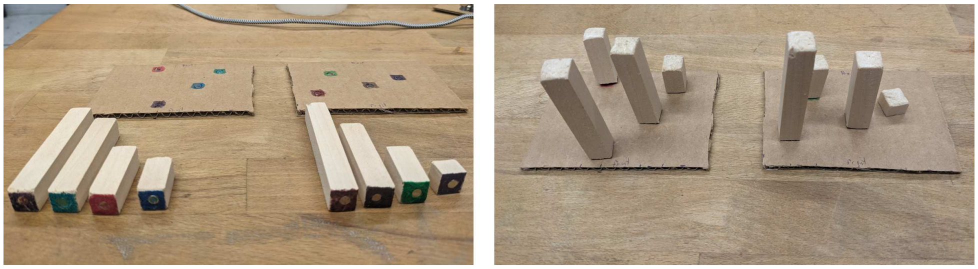


Fig. 3.7: Training example scenes: The cardboard workspace is 1:1 ratio to the 2DIY workspace with the wooden dowels being 5 mm smaller in diameter than the virtual haptic objects. (Left): Deconstructed example scenes with all objects removed from the boards. (Right): Fully constructed example scenes.

The LEGO® reconstruction medium was also replaced in favour of a more continuous reconstruction method. A laminated Cartesian sheet was placed on a metallic sheet. Borders along the laminated plate were created using hockey tape to match the workspace dimensions ratio, which is 1.75 times larger compared to the virtual environment. Objects are represented using wooden dowels affixed to magnetic bases, as seen in Fig. 3.8. The X and Z positions of objects are determined based on the magnet's location on the metallic board, whereas the vertical positions of objects are determined by the placement of hair clips along the dowels. This method of scene reconstruction gives the participants a smooth object placement along all three axes.

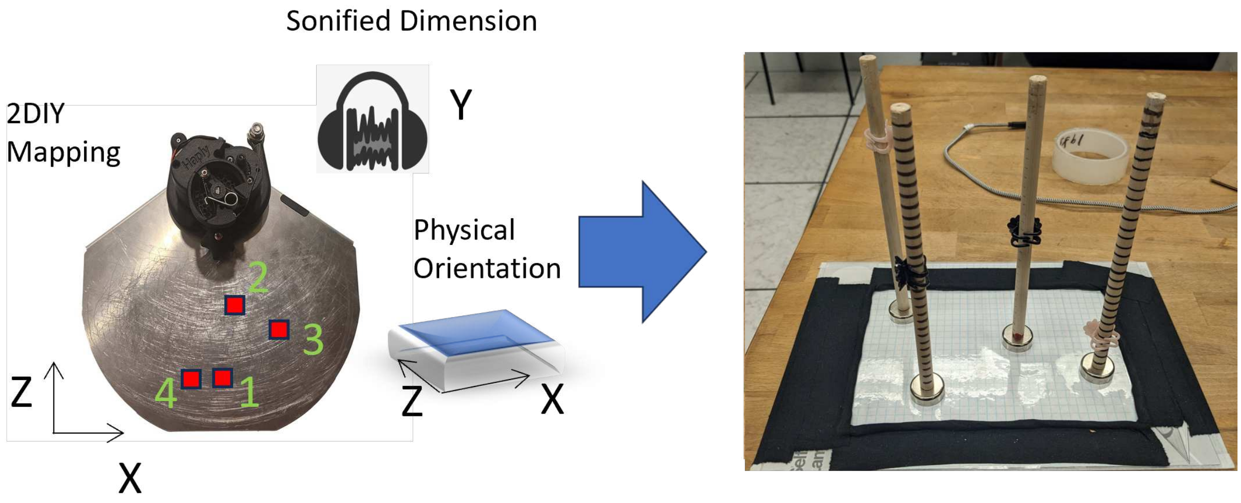


Fig. 3.8: Scene reconstruction method: Hair clips clamped onto marked wooden dowels represent the object's position in the Y-dimension, while the magnetic bases on a grid represent the object's position in the X and Z dimensions during the scene reconstruction phase as estimated by the user.

Modifications within the virtual space were also made in order to better match the physical space. The magnet bases cover more surface area than the previously used LEGO® blocks. As a result, the sizes of the virtual objects were increased to match. To improve the performance of the haptic renderings, every scene was also reprogrammed to render forces from scratch instead of the previously used physics engine, Fisica. Using Fisica, the cursor within the haptic space had a specific size, which helped with the discoverability of objects within the scene. Since stepping away from the physics engine, the cursor now occupies a single pixel, which increases the task's difficulty as the window to interact with objects has drastically decreased. The object sizes were increased by 40% of their original size leaving the haptic objects to cover a diameter of 1.5 cm within the workspace. The experimental procedure, however, remained unchanged.

Each perspective started with a training phase, followed by three parallel exploration trials and ending with a sequential exploration trial. Despite the many improvements applied to the experimental process, this version of the experiment never came to fruition for logistical reasons. Running a study over multiple days, considering the number of participants required to generate statistically significant data, required too much time and resources to accomplish within a reasonable amount of time. As a result, a third iteration of the study was created.

3.3.3.2 Third Iteration

The third iteration of this study, compared the gold standard, the physical touch of a 3D model, to that of the 2DIY top-down view of Fig. 3.3. This greatly simplified the study and could easily be executed within a single session. Physical models are created using wooden dowels glued onto a cardboard plank. The cardboard matches the size of the 2DIY workspace, and the wooden dowels are chosen to match the virtual objects' sizes as closely as possible, as seen in Fig. 3.9

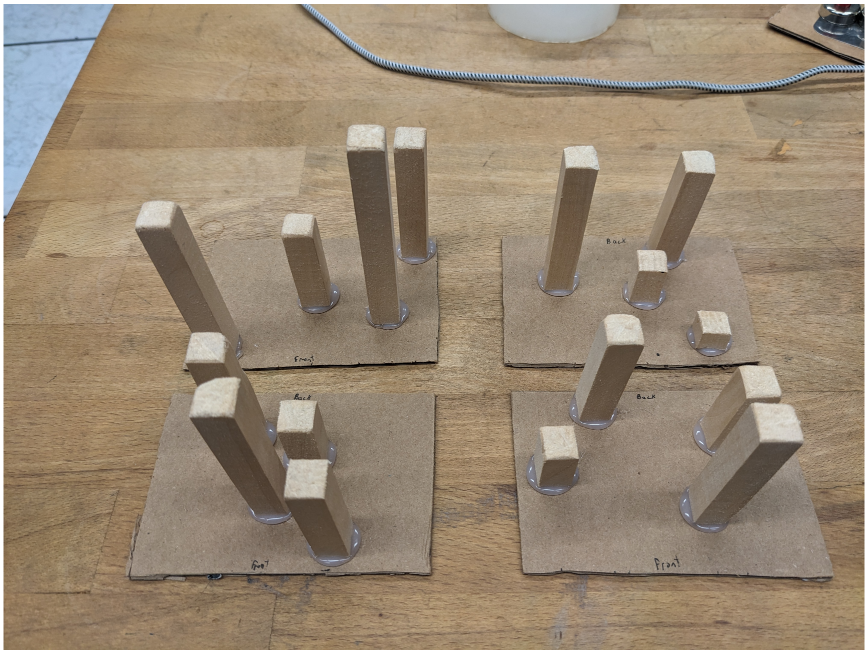


Fig. 3.9: Physical scene representation which is created to match the 2DIY workspace 1:1.

Physical touch requires no training or prior knowledge of the task, as a member of the BLV community can be expected to know how to interact with a physical model due to their day-to-day interactions. The top-down perspective was chosen as the single condition for comparison, as the literature suggests that this is likely the best way to map 3D space to a 2D medium. This is further supported by empirical observation from the pilot study, although not enough data was collected to make any claims. This new version of the study aimed to compare and quantify the performance of scene reconstruction using our approximation of a 3D space through haptics and sound against the actual 3D space of a physical model. Unlike the previous iteration from Section 3.3.3.1 that only used a physical model for training, this version also used them for the reconstruction trials. Later studies could be conducted to compare other perspectives to the gold standard. However, this version of the study, along with any other attempt to collect quantitative data, was not able to be pursued due to the device's technical limitations. Details on the 2DIY's technical constraints can be found in Appendix A.

3.3.4 2D Representation

In an attempt to reduce the mental load on participants, the dimensionality of the scenes within the study was reduced to a purely two dimensional representation. As outlined in the pilot study in Section 3.3.3, representing 3D space through two dimensional force-feedback along with pitch audio was mentally taxing on participants. Instead, the study shifted toward a heuristic approach, focusing on how participants interpreted real photographs rather than abstract points in space. More specifically, given the tools at our disposal, how would a blind participant interact with this system if it were available to them now, and how would a change in perspective representation affect a BLV individual's interpretation of a photograph? As outlined in Appendix A, the 2DIY's precision does not translate well to a quantitative study and as such a qualitative approach was taken.

Unlike previous iterations of the study that focused on synthetic environments, this version focused on real-life photographs that are extracted and processed as described in Section 3.2. The study aimed to explore the trade-offs and preferences of transforming a vision-based front-view task, such as exploring photographs, into a spatial depth-based task for the BLV community. To do so, participants' responses and preferences were compared based on their experience exploring photographs from the front-facing or top-down perspective. This version of the study had many of the same properties as the previous iteration, as described in

Section 3.3.3.1 and Section 3.3.3.2. The 2DIY was used for exploration along two dimensions, as seen in Fig. 3.10 while the size and properties of the workspace and objects within the haptic virtual environment remained unchanged from the previous iteration. However, in order to simplify the study, the third sonified dimension, the Y dimension in Fig. 3.3 and the Z dimension in Figure 3.4, was replaced with text-to-speech (TTS). Given the advanced age of many of the participants, sonofying the third dimension was not always obvious to them as described in the pilot study which is further supported by previous literature [48, 46]. The TTS in question corresponded to the object tag as detected by the object detection algorithm. This way, when the participants interacted with an object, that object's identity was announced automatically for them to hear. We determined that including the sonified third dimension using earcons required too much mental load for users to easily interpret, and it would, therefore, be more informative to only provide the names of objects during image exploration. Although the loss of three-dimensional representation may reduce the richness of information that could be provided, TTS, along with two-dimensional exploration using the 2DIY, still allowed blind users the ability to explore a virtual photo and create a mental representation of its contents. The TTS were generated through the Amazon TTS client and saved as MP3 files. The MP3 files were then read out by the Processing script locally every time a user interacts with a specified object corresponding to that TTS output. This study will compare the effects of perspective on the interpretation of photographs while exploring virtual graphics from a front-facing perspective (the camera perspective) versus a top-down perspective (from above, a bird's-eye view).

During the course of the study, four graphics were chosen to be presented to the users, which can be seen in Fig. 3.11. All scenes contain a variety of different objects, all with their own unique context. Two scenes are indoor, and two are outdoor to limit any bias to one setting or another. For any one participant, two graphics were presented from one perspective, and the two other graphics were presented in the other perspective. No participants have seen the same figure twice. Based on the user's experience, they were asked to describe the scene and answer a series of questions to determine which perspective they enjoyed and how the perspective used affected their perception of the scene.

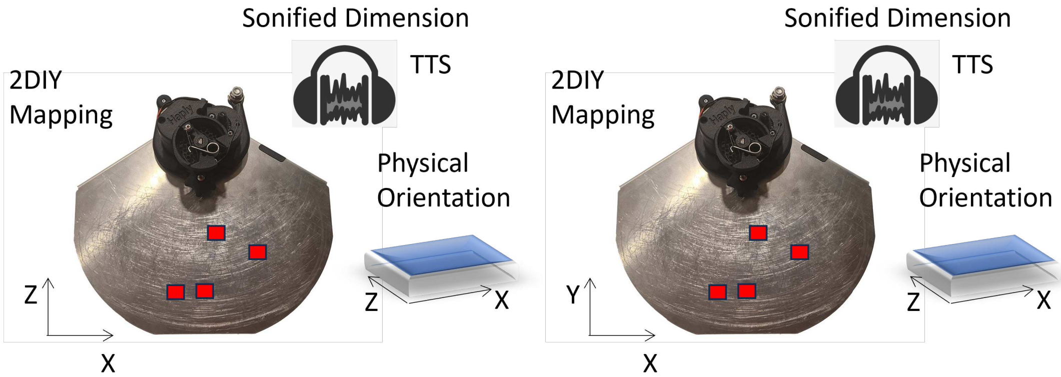


Fig. 3.10: 2DIY perspective and orientation used for 2D representation. Left is the top-down perspective, with the 2DIY exploration being along the Z-X plane. Right is the front-facing perspective, with the 2DIY exploration being along the Y-Z plane. Both machines are positioned parallel to the table.

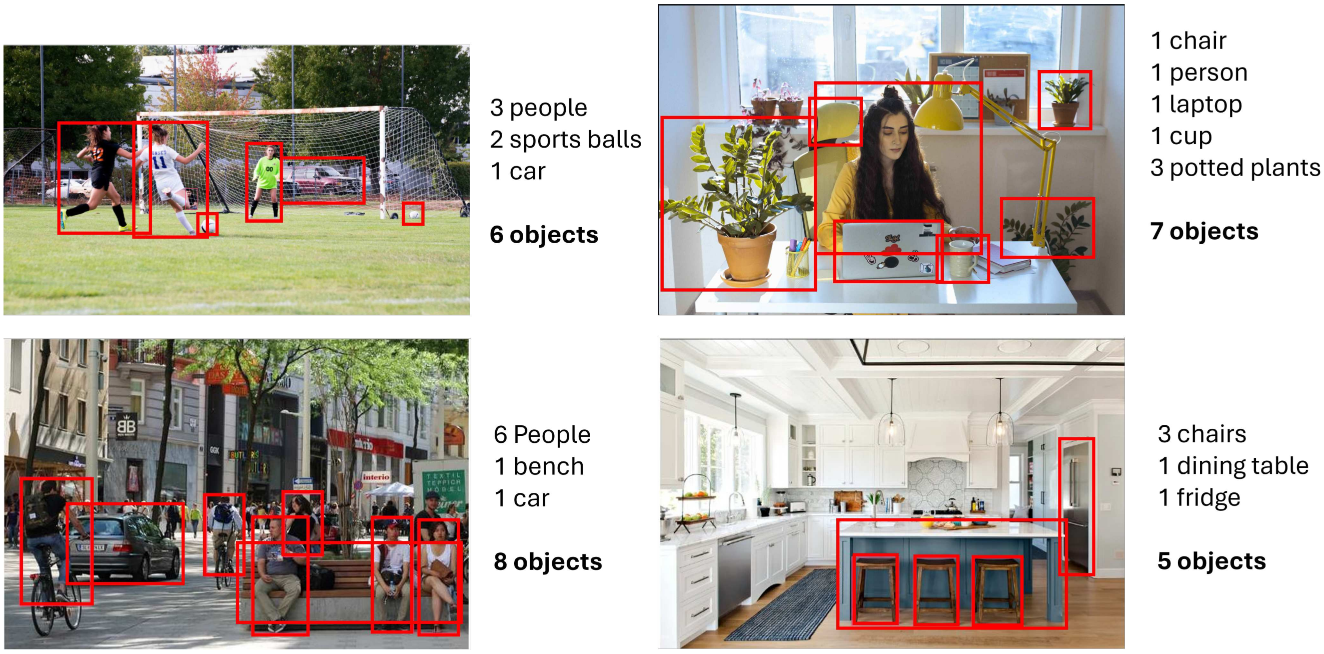


Fig. 3.11: Scenes used throughout the experiment.

Chapter 4

Interpretation of Photographs and Perspective for the BLV Community

Preface

This chapter covers the implementation of the study design from Section 3.3.4. It discusses the participant demographic, the methodology of the study's implementation, and finally, the results and discussion of the findings.

Author's Contribution

Design considerations for presenting photographs from a top-down perspective using a 2-DoF device based on data collection and empirical observation.

4.1 Introduction

Study designs were made and iterated upon to better understand the effects of perspective on photograph comprehension for the BLV community. Initially, the studies focused on participants' interpreting the specific location of objects within synthetically made 3D scenes using a 2-DoF device. These iterations were meant to quantify the effects of spatial understanding when exploring photographs from different perspectives. Eventually, this evolved into a qualitative study using real-life photographs for an ecological approach aimed to uncover the BLV community's preferences when presented with perspective. This final design, as described in Section 3.3.4, involved participants interpreting two online graphics from two perspectives, followed by a semi-structured interview.

4.2 Participants

Seven members of the BLV community were recruited through direct email and word of mouth. Since a number of participants were either anglophone or francophone, the experiment was conducted in both French and English to suit their needs. The mean age of participants was 55 years old with nearly equal gender representation (3 M and 4 F). One participant was congenitally blind, one was early blind, and the five others were late blind. Two of the late blind individuals still retained between three and ten percent of their vision. The majority of participants are well acquainted with computers and use them daily, with only two people relying solely on their cellphones and having not used a computer for over a year. The study was approved by the McGill Research Ethics Board under REB file #21-10-031. Participants were compensated CAD 15/hour for their time.

4.3 Methodology

4.3.1 Procedure

Over the course of the experiment, video and audio were recorded for future analysis. The experiment was conducted either on campus within lab facilities or at the participant's domicile, whichever was more convenient for the participant. Once participants agreed to the experiment, a pretest questionnaire is administered as seen in the Appendix A 5 that covered basic personal questions and details about the nature of their blindness. A simple physical model of a scene is provided as shown in Fig. 4.1 as the first scene introduction to assist participants in conceptualizing the task, as the researcher provides a detailed verbal description of the perspective.

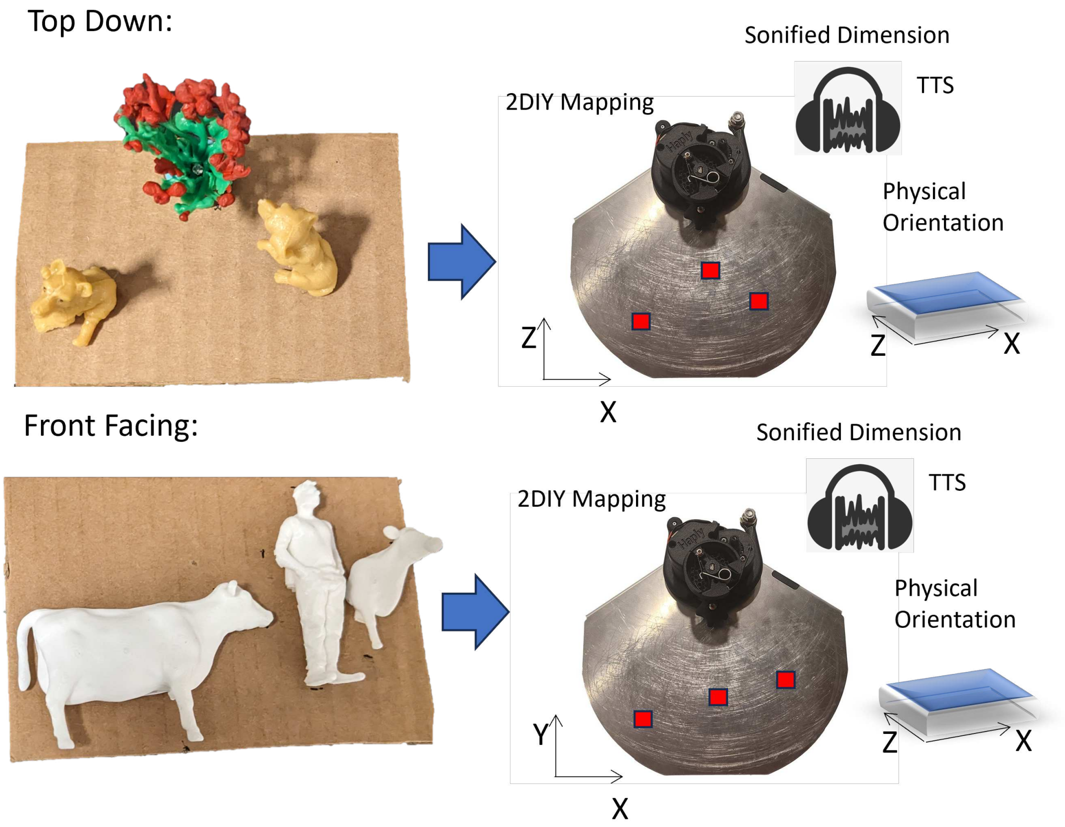


Fig. 4.1: Physical models used to help blind users conceptualize the scene with their corresponding 2DIY representation to the right.

After describing the scene that corresponds to the physical model, the researcher ensures that any confusion is cleared and proceeds to guide the participant through the haptic rendering that matches the physical model. The participant can then explore the virtual scene along with the physical model simultaneously. Once the individual is satisfied and the training is complete, the experimental trials begin. One of the graphics provided in Fig. 3.11 is rendered on the 2DIY, and the participants are informed of the number of objects present within the scene prior to exploration. Once the user has explored all the objects and is satisfied with their exploration, the researcher asks them a series of questions as provided in Appendix A 5. After gathering information from the participant, the true description of the scene is revealed, followed by more questions from the researcher. Following completion of this step, a new scene is chosen for the same perspective and the process is repeated. Once two scenes are explored, another series of questions regarding that particular perspective is administered, which then concludes that perspective. The training, testing, and question process is then repeated again for the second perspective using two different scenes in Fig. 3.11. The order of presentation of perspectives is counterbalanced across participants and each scene is counterbalanced across each perspective. After all four scenes have been presented, two in each perspective, the final questionnaire, in is administered to obtain their final thoughts, feedback, and preferences from each perspective.

4.4 Results

4.4.1 Interview Results and Observations

Throughout the experiment, participants were tasked with exploring the haptic space using the 2DIY and describing the scene back to the researcher, followed by a series of questions. When asked how confident the participants were with their interpretation of the scene (from a scale of one to five), it was found that they were more confident in their answers when describing a scene from a top-down view than a front-facing one (top-down: $3.91 \pm \sigma = 0.87$, front-facing: $3.54 \pm \sigma = 0.87$), although more data must be collected for statistical significance.

When asked which perspective they preferred, three out of the seven participants chose the top-down approach, two preferred the front-facing view, one saw the benefits of both strategies and could not decide, and another who, while understanding the difference between the perspectives, could not notice a change in the way they were presented. Despite this, it is noted that this participant was still able to get a general sense of where objects are located depending on the information available for each perspective (i.e., they knew the vertical position of objects when in front-facing view and not for top-down). This may suggest that they understood the task and performed as intended but the lack of difference between perspectives, such as the 2DIY position being the same as well as the normalized object shapes and haptics across both conditions, caused both perspectives to "feel" the same which led them to be unable to pick a favoured approach.

Of the two participants who chose the front-facing perspective as their preferred interpretation method, one had the best eyesight of the cohort (10 % vision) as well as being blind for the shortest time among the participants. With that in mind, it is no surprise that this participant is more comfortable with a vision-based approach to photographs. Meanwhile, the second participant, who preferred the front-facing perspective, claims that the top-down view of a photograph feels unnatural and clashes with their memory of what photographs represent.

The three participants who preferred the top-down view of photographs claim that it "makes sense" as well as provides a strong spatial relationship between the objects within the scene. The only congenitally blind participant was among those who preferred this perspective. This aligns with our hypothesis that people without any visual experience would benefit from this interpretation of photographs, although more tests are required to make any claims. Interestingly, the other blind individual who still retains sight (3 % vision) also preferred this method. This suggests that this perspective of photographs may even be useful for individuals who still rely on sight as well as those who are completely blind, such as the other two who preferred this method.

Finally, the individual who could not decide which perspective they preferred cited that both perspectives had merits, and depending on the task, one would be more beneficial over the other. This early blind individual agreed that a top-down perspective is ideal for spatial tasks in order to determine the position of objects within a scene. However, they also stated that a front-facing view would be better suited for a detail-oriented task, such as exploring a portrait. As such, they could not select a single perspective as they do not offer solutions to all potential photographs.

4.5 Discussion

4.5.1 Perspective

Feedback from study participants, as seen in Section 4.4.1, indicated that most of the BLV individuals tested had preferred the top-down view over the front-facing perspective when exploring photographs, especially with respect to the spatial representation of objects which supports previous literature [16]. However, given the small sample size, these results were not statistically significant. When describing scenes in Fig. 3.11, differences in how they were interpreted manifested themselves depending on the perspective used during exploration.

The extent of these variations changed from scene to scene. The kitchen photograph, for example, did not show much variability between perspectives. In both cases, users were unclear about the placement of the chairs in relation to the table, as details regarding the size of the table were missing. All participants were able to determine that the scene was taking place in a kitchen after discovering the refrigerator but continued to explore the photograph in search of other cues to validate their assumption, such as a sink, counter, or stove. Similarly, the vertical positions of objects in the scene were relatively centered across the mid-section of the photograph, making the front-facing view not particularly informative either, as the objects had similar heights. Since the detected objects were not spread out at different distances throughout the space, the top-down perspective did not offer many advantages. This outlined a weakness in the way the scenes were presented in our work. Normalizing the object shapes to a circle reduced the fidelity of the scene, drastically affected the user's ability to interpret the position of objects in relation to one another. Due to this lack of context and shape, the presented perspectives did not offer any advantage over each other for this particular scene.

Similar to the kitchen graphic, the street scene did not show much difference in interpretation between perspectives. This scene, which contained eight objects (six people, a bench, and a car), was initially daunting to users. Despite this, each participant was able to discover every object throughout the exploration phase in a matter of minutes. Although the difference in how the scene was interpreted did not drastically change from one perspective to the next, the depth and vertical position of objects did not go unnoticed by participants. In the front-facing perspective, one participant noticed the different heights of the people within the scene and surmised that the individuals who were lower within the photograph were children, while the taller people were adults. Within this perspective, it was also noted that providing shape to the objects would have greatly improved understanding of the interaction between two objects. When using the top-down perspective, users could generally identify that the car was located further away in the background. However, due to its lack of shape, users were unable to confidently determine how the people nearby were interacting with the vehicle.

The variation between the perspectives became clearer when participants were asked to interpret the soccer scene. This scene contained three people, two sports balls, and a car for a total of six objects. Participants using the front-facing perspective attempted to imagine a scene that included all of the objects present as key features, which led to significant confusion. As a result, only one out of four participants who received this scene in front-facing view managed to guess a location of where this photograph had been taken. Those who were unable to guess a location could not imagine a situation in which two balls, a car, and three people could interact and, therefore, could not come to any conclusions about the nature of this scene. Meanwhile, all participants who explored the scene from a top-down perspective were able to guess where the scene was taking place and what sport was being played. Users guessed that the scene was taking place in a field or driveway where people were playing soccer or basketball. Unlike the front-facing perspective, participants from the top-down view were able to determine that the car and second ball were not essential components of the scene and would, therefore, consider them as background pieces. Providing depth information, as seen from the top-down perspective, had a major impact on the participants' ability to interpret the scene, allowing them to parse foreground and background objects. Since the front-facing view projected all objects to the foreground of the picture, determining what was important to the scene became incredibly challenging.

Finally, the office scene, a familiar environment for all participants, had many similarities between perspectives with a few key differences. Regardless of perspective, participants were able to identify this scene as a work setting due to the laptop and nearby person. Specifically, they identified this scene as a home office due to the number of potted plants present in the photograph. Given the proximity of the chair, laptop, cup, and person, everyone also assumed that the person was sitting on the chair working on the laptop and enjoying a coffee. Users who were presented the scene from the front-facing perspective noticed the various heights of the objects within the scene but did not come to any novel conclusions compared to the other perspective, despite difference in point of view. On the other hand, participants who were provided this scene from the top-down perspective noted that the distance between the person and the chair was greater than expected and often wondered whether the person was sitting on the chair or not. However, the congenitally blind individual took this as a cue of the person's posture and correctly concluded that the person was not only sitting on the chair but also leaning forward, causing a greater gap between the person and the chair. Provided with basic information about the objects within the photograph, the top-down perspective allowed BLV individuals to infer more information about the scene compared to the front-facing view.

Based on our observations, the top-down perspective enabled participants to gain a better spatial understanding of a scene, parse background and foreground objects, and infer additional details about a photograph's contents. This aligns with previous observations and underscores its potential for the BLV community [16, 17].

In contrast, the front-facing perspective allowed users to determine the vertical positioning of objects, which provided some context of the scene but does not offer much additional information otherwise. As described by our congenitally and early blind participants, this perspective was ultimately less informative than the top-down view for blind users, especially for spatial information. However, they also mentioned that it could be more informative in a more detail oriented context such as a portrait, where the use of space and actions are no longer needed.

4.5.2 Context

As outlined by Kennedy et al. [13], context played a crucial role in a BLV individual's understanding of tactile graphics. Throughout this study, participants were provided with minimal cues: normalized objects, identical haptic effects, and an audio tag for each object. This, in turn, forced participants to rely on rudimentary data and object positioning to infer scene context.

For familiar settings, like the kitchen and home office, participants were able to reliably deduce context in both perspectives. However, in unfamiliar environments, such as the soccer scene, those using the front-facing perspective were unable to infer context and struggled to describe the scene, whereas participants using the top-down perspective showed far more success. This highlighted the importance of distinguishing between foreground and background objects, particularly when participants could not rely on past experiences to bridge any knowledge gap within the tactile scene. When provided with this spatial information, participants were able to identify key foreground objects while simultaneously using background elements to build additional context.

Participants frequently relied on contextual reasoning to fill in gaps left by incomplete or limited sensory information provided by the system. In the office scene, participants commented on the absence of a table within the photo and were often found exploring the haptic space in an attempt to find it. When asked to describe the scene, all four participants who explored this scene from a top-down view and one front-facing user still included a table in their description, despite none being presented. This may suggest that the top-down perspective provided greater context cues, allowing participants to infer more information about the scene.

The importance of context was especially evident in the kitchen scene. All participants were able to determine that the scene was taking place in a kitchen after discovering the refrigerator but continued to explore the photograph in search of other cues to validate their assumption. The object detection algorithm also classified the center island table as a dining table, which often led users to wonder why there were only three chairs, as dining tables usually came in pairs of two to six chairs. Most of the participants believed that the number of chairs present had significant meaning, but they lacked the necessary information to reach any conclusions. Once it was verbally revealed that the dining table was an island table, participants were pleased to have found a solution for the odd number of chairs in the scene.

Similarly, in the street scene, context significantly impacted participants' interpretations. Given the high person count within the scene, users had difficulty knowing what people were doing without having the actions explicitly stated. As such, objects within the scene appeared to be grouped based on proximity. The two people on the left of the photograph were generally considered to be inside or interacting with the car on the left. Meanwhile, on the right-hand side of the photo, the two people sitting next to each other on the bench were paired and were assumed to be sitting on the bench. All other people in the center of the graphic were assumed to be involved with the people near the bench, with the scene typically described as a social setting in a park or a group of people waiting for transport. All participants commented that the lack of knowledge regarding the people's actions greatly diminished their ability to generate a mental representation of the scene.

Contextual confusion peaked in the soccer scene. The soccer balls being identified as a "sports ball" made recognizing the context difficult, as there were many sports that this scene could describe. The fact that there were two sports balls present also confused the participants, as sports usually have one ball in play at a time. Furthermore, the presence of the car in the background often misled the BLV individuals into believing that it was an important aspect of the scene when it was simply parked in the background. Participants exploring the front-facing perspective particularly struggled, attempting to imagine a cohesive scenario that included all visible objects as key features. By contrast, the participants exploring the top-down perspective quickly identified the car and extra ball as background elements, and thus more accurately inferred the activity taking place.

These examples illustrated clearly that context was an essential factor for BLV users when interpreting scenes. The observations from this study supported the literature which emphasized that, even with limited tactile information, contextually-rich cues significantly aided comprehension[13]. Participants' feedback further underscored the importance of either providing richer tactile cues, such as shapes, edges, and textures, or a concise verbal synopsis prior to exploration to reduce the ambiguity and enhance overall scene understanding.

4.5.3 Space & Navigation

Participants demonstrated a robust ability to mentally map and navigate the presented scenes. This capability aligned closely with O&M training methods and literature, emphasizing landmark-based and allocentric navigation strategies commonly used within the BLV community[14, 15]. Many participants intuitively traced the rectangular workspace boundaries before systematically moving inward to locate landmarks and objects.

In the street scene, despite initially finding the high object count daunting, participants quickly grouped objects based on proximity into logical clusters, such as two people near the car, two individuals seated on the bench, and others centrally placed. The depth differentiation inherent in the top-down view facilitated efficient spatial mapping, allowing users to quickly distinguish foreground from background elements. Conversely, participants exploring the front-facing perspective were able to determine the vertical positioning of objects and thus could infer people sitting on the bench, however faced difficulty to infer much else spatially due to the depth compression.

The soccer scene also highlighted the spatial advantages of the top-down view. Participants using this point of view recognized that one of the balls was located farther back, thus dismissing it as likely out-of-play. This depth recognition allowed them to correctly infer that the primary activity involved three people interacting around one ball, indicative of a sport like soccer or basketball. In contrast, participants exploring the front-facing perspective treated all objects as being in the foreground, significantly complicating their spatial understanding of what was transpiring within the scene.

The better spatial understanding afforded by the top-down view supports existing literature, which suggests that this perspective mirrors a blind person's navigation strategy, as outlined by Szubielska et al. [11]. Participant comments further validated this finding, with many noting that the top-down perspective "just makes sense", meaning that it already aligns with their perspective of the world. Conversely, participants who favored the front-facing perspective appeared to adopt a vision-based approach, reconciling it with their visual expectations of photographs. These differing strategies echo patterns in spatial perception observed between blind and sighted children, as described by Martolini et al. [19].

Despite this, both perspectives encountered limitations due to uniform object shapes and insufficient spatial granularity. Normalizing all object shapes into identical circles drastically reduced participants' ability to accurately interpret positions and relationships between objects. In the kitchen, for instance, participants struggled to align chairs correctly to the island table due to the lack of object edges and contours, as there was no clear indication of where the table began and ended. Similarly, in the office scene, participants found it difficult to differentiate whether potted plants were placed on the ground or on a surface like a table. Users explicitly requested clearer shapes, defined edges, and textured surfaces, to place themselves into the space allowing them to feel the scene as well as where the objects are located within it.

A follow-up study is being designed that incorporates lessons learned from the observations made in our work. The new study focuses on the top-down perspective and aims to create an idealized experience for the BLV community that is beyond what is currently available. This includes the addition of object shapes, more scene information such as the presence of grass or concrete, and a greater emphasis on haptic feedback. As demonstrated in Zhang et al. [38], FFDs can be used to great effect to convey detailed haptic information that would otherwise be difficult to transmit. The work outlined in our research does not yet fully leverage the capabilities of the 2DIY, and we believe that improving the haptic effects within the scene can create a more compelling experience. In other words, rather than applying the same spring, texture, and damping force to every object as we did in our work, each object would have forces that mimic its real sensation, such as a jagged texture for a concrete road or textured damping to represent a person. As discussed previously, a large drawback of using normalized object sizes is the loss of certain spatial cues and context. For example, participants exploring the street scene were unclear about the people's interaction with the car. Had the shape of the people and car been available, the context between the objects would have been much clearer. As such, the future study aims to rectify this problem by providing shape to the objects for a participant to explore. A comment that was repeatedly made by participants is the lack of context of the scene's surroundings. Although that was outside the scope of our work, it is nevertheless an essential aspect of a photograph. In the future study, this information will be presented as haptic feedback through the 2DIY to allow users to feel the surrounding space within the photograph, not only the objects. Additionally, a description of the scene will be provided to the participants prior to the haptic representation in order to greatly increase their understanding of its contents by giving a basis on which to build their assumptions, similar to the work done using tactile displays by Kennedy et al. [13] Altogether, the future work will create a system that leverages the strengths of the spatial representation of the top-down perspective while simultaneously amending the discovered pitfalls.

Chapter 5

Conclusion

Depicting a photograph to the BLV community presents a challenge for designers and researchers. Previous work relied on strategies such as encoded pixel values to a sonification or created tactile feedback through braille displays or tactile graphics to portray graphics. However, the software solutions did not take advantage of the sense of touch, while the hardware solutions were not affordable. The current standard relied upon by the BLV community, alt-text and TTS, simply provides a synopsis of the graphic's content, which delivers a relatively limited representation of the scene. This thesis explored a novel approach to photographs by transforming previously front-facing graphics into a top-down perspective, allowing users to explore the scene through touch and audio.

Designing a 3D space for a 2-DoF device required using sonification to compensate for the missing dimension. Various audio solutions were considered due to the advanced age of a large subset of the BLV population and their associated hearing difficulties. Earcons using pitch to deliver the third dimension were chosen for their rapid information delivery and ability to be perceived by the elderly population.

A pipeline was created through the IMAGE architecture that allowed users to extract graphics from websites through their web browser and discover objects within the scene. Four depth extraction algorithms were considered to approximate the distance of objects within photographs, and one was ultimately selected thanks to its output quality, low resource consumption, and rapid processing time. Using the information gathered from this pipeline, users are able to determine the location of objects within a photograph in three dimensions.

Multiple iterations of the study were designed to present this three-dimensional space using a two-dimensional device. The first study design exposed the participants to three different perspectives and 2DIY orientations. Users were tasked to explore the 2DIY workspace and reconstruct the object placement within the scene using LEGO®. Preliminary testing on three blind participants revealed key flaws in the initial study design. The second iteration of the study aimed to rectify these flaws by running the study over two days, reducing the number of perspectives to two, improving training with physical props, and changing the LEGO® reconstruction to a more continuous medium using magnets and dowels. For logistical reasons, the study could not be conducted. A third iteration was developed that aimed to simplify the study even further by comparing a single perspective with 3D models. This allowed experiments to be conducted in a short amount of time within a single session. Technical flaws in the 2DIY and the device's lack of precision made these quantitative studies impossible to conduct. To cope with these challenges, a final version of the study was created.

The resulting qualitative study compared two perspectives on real-life photographs in a semi-structured style interview and aimed to determine the effects and preferences of perspective for the BLV community. The third dimension, using sonification, was replaced with TTS to identify the objects, reducing the complexity of the tasks. The study was conducted on seven blind individuals and the results showed that most preferred the top-down perspective when exploring photographs. Participants were also found to be more confident in their responses when asked to describe the scene. The top-down perspective allowed users to parse foreground and background objects, while all objects in the front-facing perspective appeared to be in the foreground, causing confusion. From the top-down perspective, more inferences could be made about the scene, adding additional detail to its interpretation over the front-facing view.

This thesis showed that exploring photographs from a top-down perspective was beneficial and could be used as a superior alternative to the traditional front-facing perspective for the BLV population. Our work outlined numerous design considerations to be made when presenting a scene from the top-down perspective and offered suggestions for future studies. Our contributions also provided researchers with the means to convert a front-facing photograph into an approximate top-down representation through the IMAGE architecture. This, in turn, allows researchers to create their own experiences and, based on our findings, make informed decisions on the BLV community's preferences when designing them.

However, certain limitations persist. Currently, the system is unable to generate complex shapes and is limited to representing objects as circles, depriving users of potentially useful details. Additionally, background information, such as the floor composition, was not included in this experience, greatly limiting the contents of what is presented in the virtual photograph. The lack of shape of objects and contextual cues did not provide enough details for a full user experience and required more development to become a deployable system.

Future studies should be conducted to validate whether increasing the fidelity of objects within the scene has the intended effect on scene understanding. Based on the results of this study, features such as shape, as well as object and background-specific textures would greatly enhance the user experience. To accomplish this, it would be useful to create this experience using a 3D scene to gauge the community's outlook on an ideal case and test the limits of the top-down representation of photos.

References

[1] L. Pelchen, “Internet usage statistics in 2024.” https://www.forbes.com/home-improvement/internet/internet-statistics, March 2024. [Accessed 29-07-2024].

[2] S. Shethia and A. Techatassanasoontorn, “Experiences of people with visual impairments in accessing online information and services: A systematic literature review,” *Pacific Asia Journal of the Association for Information Systems*, pp. 39–66, January 2019.

[3] WebAIM, “The 2024 report on the accessibility of the top 1,000,000 home pages.” https://webaim.org/projects/million/#alttext, March 2024. [Accessed 11-05-2024].

[4] UserWay, “Empowering independence: enhancing web accessibility for the blind.” https://userway.org/blog/web-accessibility-for-blind-users. [Accessed 08-07-2024].

[5] R. Absar and C. Guastavino, “Usability of non-speech sounds in user interfaces,” *Proceedings of the 14th International Conference on Auditory Display (ICAD ’08)*, June 2008.

[6] K. Pakėnaitė, P. Nedelev, E. Kamperou, M. J. Proulx, and P. M. Hall, “Communicating photograph content through tactile images to people with visual impairments,” *Frontiers in Computer Science*, vol. 3, 2022.

[7] M. Mukhiddinov and S.-Y. Kim, “A systematic literature review on the automatic creation of tactile graphics for the blind and visually impaired,” *Processes*, vol. 9, no. 10, 2021.

[8] J. A. Daniel Shin, “For blind people, technology can offer a way to perceive images through touch.” https://www.marketplace.org/shows/marketplace-tech/printer-tactile-images-blind-people/, October 2022. [Accessed 23-07-2024].

[9] M. Butler, L. M. Holloway, S. Reinders, C. Goncu, and K. Marriott, “Technology developments in touch-based accessible graphics: A systematic review of research 2010-2020,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, (New York, NY, USA), Association for Computing Machinery, 2021.

[10] J. M. Kennedy, *To Touch and to Picture the World*. New Haven, CT: Yale University Press, 1st ed., 1993.

[11] M. Szubielska, E. Niestorowicz, and B. Marek, "Drawing without eyesight. Evidence from congenitally blind learners," *Roczniki Psychologiczne*, vol. 19, no. 4, pp. 681–700, 2016.

[12] M. N. John K. Gilbert, Miriam Reiner, *Visualization: Theory and Practice in Science Education*. Springer Dordrecht, 2008.

[13] J. M. Kennedy, "Blind people and outline drawings," in *Drawing & the Blind: Pictures to Touch*, ch. 3, pp. 57–94, New Haven, CT: Yale University Press, 1st ed., 1993.

[14] T. Schwartz, "Orientation and Mobility: A Guide for the Visually Impaired." https://lifeafterblindness.com/orientation-and-mobility-a-guide-for-the-visually-impaired/, February 2024. [Accessed 23-04-2024].

[15] S. Ungar, "Cognitive mapping without visual experience," *Cognitive Mapping*, 2018.

[16] J. M. Kennedy, "Perspective," in *Drawing & the Blind: Pictures to Touch*, ch. 6, pp. 180–215, New Haven, CT: Yale University Press, 1st ed., 1993.

[17] J. M. Kennedy, "Drawings by the blind," in *Drawing & the Blind: Pictures to Touch*, ch. 4, pp. 95–126, New Haven, CT: Yale University Press, 1st ed., 1993.

[18] Y. Eriksson, "How to make tactile pictures understandable to the blind reader," in *in Proceedings of the 65th IFLA Council and General Conference*, 2010.

[19] C. Martolini, G. Cappagli, E. Saligari, M. Gori, and S. Signorini, "Allocentric spatial perception through vision and touch in sighted and blind children," *Journal of Experimental Child Psychology*, vol. 210, p. 105195, November 2021.

[20] A. Bhattacharjee, A. J. Ye, J. A. Lisak, M. G. Vargas, and D. Goldreich, "Vibrotactile masking experiments reveal accelerated somatosensory processing in congenitally blind braille readers," *Journal of Neuroscience*, vol. 30, no. 43, pp. 14288–14298, 2010.

[21] C. L. Reed and M. Ziat, "Haptic perception: From the skin to the brain," in *Reference Module in Neuroscience and Biobehavioral Psychology*, Elsevier, 2018.

[22] L. M. Briana Fraser, "AI generated Alt Text," in *Accessibility Handbook for Teaching and Learning*, Langara College, August 2023.

[23] H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell, “Understanding blind people’s experiences with computer-generated captions of social media images,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI ’17, (New York, NY, USA), p. 5988–5999, Association for Computing Machinery, 2017.

[24] E. Anderson, “Pixel composition: Converting images to music,” in *WWU Honors College Senior Projects*, 2020.

[25] A. P. Jorge, “The music of paintings: a rhythmic perspective,” in *12th Generative Art Conference*, December 2009.

[26] M. Banf and V. Blanz, “Sonification of images for the visually impaired using a multi-level approach,” *ACM International Conference Proceeding Series*, pp. 162–169, March 2013.

[27] G. Dubus and R. Bresin, “A systematic review of mapping strategies for the sonification of physical quantities,” *PLOS ONE*, vol. 8, pp. 1–28, December 2013.

[28] N. R. C. (US), “The demography of blind and visually impaired pedestrians,” in *Working Group on Mobility Aids for the Visually Impaired and Blind.*, ch. 2, Electronic Travel AIDS: New Directions for Research. Washington (DC): National Academies Press (US), 1986.

[29] K. Pakenaite, E. Kamperou, M. J. Proulx, A. Sharma, and P. Hall, “Pic2tac: Creating accessible tactile images using semantic information from photographs,” in *Proceedings of the Eighteenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI ’24, (New York, NY, USA), Association for Computing Machinery, 2024.

[30] L. Zeng, M. Miao, and G. Weber, “Interactive Audio-haptic Map Explorer on a Tactile Display,” *Interacting with Computers*, vol. 27, pp. 413–429, February 2014.

[31] R. Klatzky, S. Lederman, and C. Reed, “There’s more to touch than meets the eye: The salience of object attributes for haptics with and without vision,” *Journal of Experimental Psychology: General*, vol. 116, p. 356, December 1987.

[32] NationalBraillePress, “The need for braille.” Https://www.nbp.org/ic/nbp/about/aboutbraille/needforbraille.html, 2024. [Accessed: 2024-05-10].

[33] Do-It, “Are touch screens accessible?.” https://www.washington.edu/doit/are-touch-screens-accessible, May 2022. [Accessed 08-07-2024].

[34] G. Jansson, “Can a haptic force feedback display provide visually impaired people with useful information about texture roughness and 3d form of virtual objects,” 1998.

[35] F. Carneiro, M. Quintas, P. Abreu, and M. Restivo, “Design and test of a 1 dof haptic device for online experimentation,” *International Journal of Online Engineering (iJOE)*, vol. 12, p. 55, April 2016.

[36] G. Jansson, H. Petrie, C. Colwell, D. Kornbrot, J. Fänger, H. König, K. Billberger, A. Hardwick, and S. Furner, “Haptic virtual environments for blind people: Exploratory experiments with two devices,” *International Journal of Virtual Reality - IJVR*, vol. 4, January 1999.

[37] D. Lareau and J. Lang, “Haptic rendering of photographs,” in *2012 IEEE International Workshop on Haptic Audio Visual Environments and Games (HAVE 2012) Proceedings*, pp. 107–112, October 2012.

[38] T. Zhang, B. S. Duerstock, and J. P. Wachs, “Multimodal perception of histological images for persons who are blind or visually impaired,” *ACM Trans. Access. Comput.*, vol. 9, January 2017.

[39] R. I. Tivadar, B. Franceschiello, A. Minier, and M. M. Murray, “Learning and navigating digitally rendered haptic spatial layouts,” *Nature News*, December 16 2023.

[40] J. Regimbal, J. R. Blum, and J. R. Cooperstock, “IMAGE: A deployment framework for creating multimodal experiences of web graphics,” in *Proceedings of the 19th International Web for All Conference*, pp. 1–5, 2022.

[41] Y. Ji, Y. Li, X. Sun, S. Yan, and N. Guo, “Stereo matching algorithm based on binocular vision,” in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, pp. 843–847, 2020.

[42] W. Yin, Y. Liu, and C. Shen, “Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[43] A. Masoumian, H. A. Rashwan, S. Abdulwahab, J. Cristiano, M. S. Asif, and D. Puig, “Gcndepth: Self-supervised monocular depth estimation based on graph convolutional network,” *Neurocomputing*, 2022.

[44] S. M. H. Miangoleh, S. Dille, L. Mai, S. Paris, and Y. Aksoy, “Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging,” *Proc. CVPR*, 2021.

[45] A. Masoumian, D. Marei, S. Abdulwahab, J. Cristiano, D. Puig, and H. Rashwan, *Absolute Distance Prediction Based on Deep Learning Object Detection and Monocular Depth Estimation Models*. Open Access, October 2021.

[46] S. Merchel and M. E. Altinsoy, “Psychophysical comparison of the auditory and tactile perception: a survey,” *Journal on Multimodal User Interfaces*, 2020.

[47] G. Dubus and R. Bresin, “A systematic review of mapping strategies for the sonification of physical quantities,” *PLOS ONE*, vol. 8, pp. 1–28, December 2013.

[48] S. Merchel and M. E. Altınsoy, “Psychophysical comparison of the auditory and tactile perception: A survey,” *Journal on Multimodal User Interfaces*, vol. 14, July 2020.

[49] S. J. J. Hopwood, *Science Music*. New York, The Macmillan Company; Cambridge, Eng., The University Press, 1937.

[50] W. Thalheimer, “Spacing learning events over time: What the research says.” http://www.work-learning.com/catalog/, February 2006. [Accessed 12-05-2024].

Appendix A

2DIY Testing results

Certain technical limitations of the 2DIY arose throughout the testing and development of our study, which prevented the project from moving forward with a quantitative focus. One such limitation was a result of the 2DIY having an offset within its haptic environment. As illustrated in Fig. 5.1, the location of a virtual haptic object produced by the 2DIY changed depending on the direction in which the user approached it.

This offset first went unnoticed, as the visual rendering of the haptic scene did not have this offset and was only discovered by doing blindfolded testing. Upon further investigation, it was found that the device's visualizations did not match its physical outputs. This meant that visual debugging of the scene was not possible, as the device was unable to accurately keep track of its physical location. In order to diagnose this problem, a script was made containing a series of ten boxes of the same size located at different points within the workspace as seen in Fig. 5.3. Once within the boxes, the end-effector would receive a light texture vibration to help indicate that you are located within the square, and whenever the end-effector crosses an object border (incoming or outgoing), there would also be a bell sound to indicate that they have crossed a boundary. Using this script, the end-effector would be passed over each object from different incident directions. Each time a border was crossed, the location of the object was physically marked on a piece of paper located below the device. The offset was then measured as the distance between the incoming and outgoing positions of an object's border edge. This was repeated across the entire workspace on multiple 2DIYs. What was discovered is that the offsets were present across all devices, and the severity of the offset varied within the workspace. Depending on the device and workspace position, the distance between per-

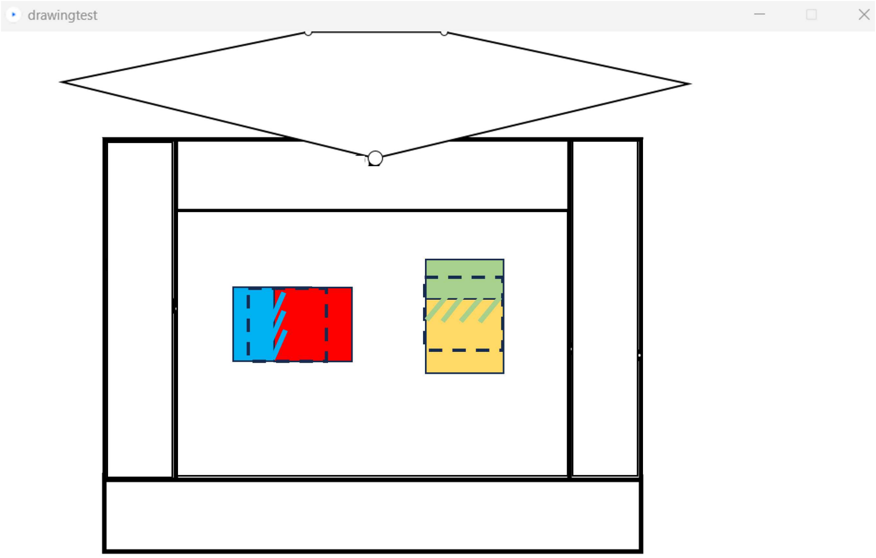


Fig. 5.1: 2DIY offset: The dashed box outline represents the object's true position. The blue indicates the haptic sensation of an object when approaching on the right, while the red indicates the location of the same haptic object when approaching from the left. The same effect is observed when approaching objects from above (yellow) and from below (green).

ceived object locations could vary from a few millimeters to a full centimeter. In hindsight, the participants in the pilot study in Section 3.3.3 were likely able to notice this offset, where they would often comment on the difficulty of finding previously discovered objects. After an internal investigation in conjunction with lab collaboration, the Haply Robotics team discovered that the elastic encoder bands were deforming while under use. The elastic deformation caused the device's physical output to mismatch the intended output, making the deformations repeatable within a device but variable across the workspace.

The fix employed by our industry partners involved replacing the pulley and elastic system which was used to apply forces from the motors to the end-effector to a gear and belt system. This method effectively guarantees that there are no deformations from the forces applied by the device, as well as drastically increases the accuracy of the machine.

However, it was quickly discovered that the forward kinematics using this new system did not match as they had initially anticipated. As seen in Fig. 5.2, drawing a straight line with the end-effector in the physical world did not result in a straight line as interpreted by the 2DIY, instead it drew a distorted wave. In addition to the mismatching kinematics, the workspace located closest to the motors gave the impression that friction forces were being applied. Haply Robotics addressed these issues by changing the gear ratio of the 2DIY, which also addressed the worst of the friction problem near the motors.

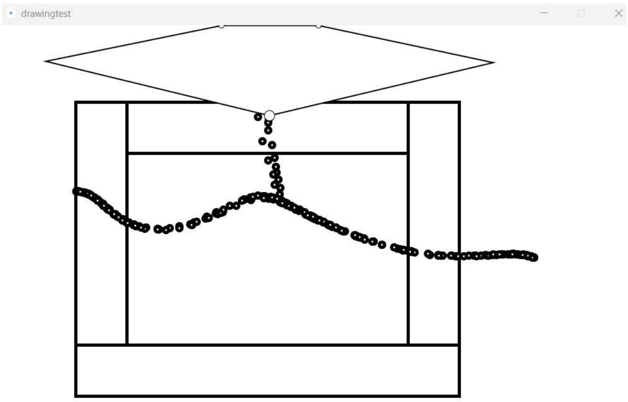
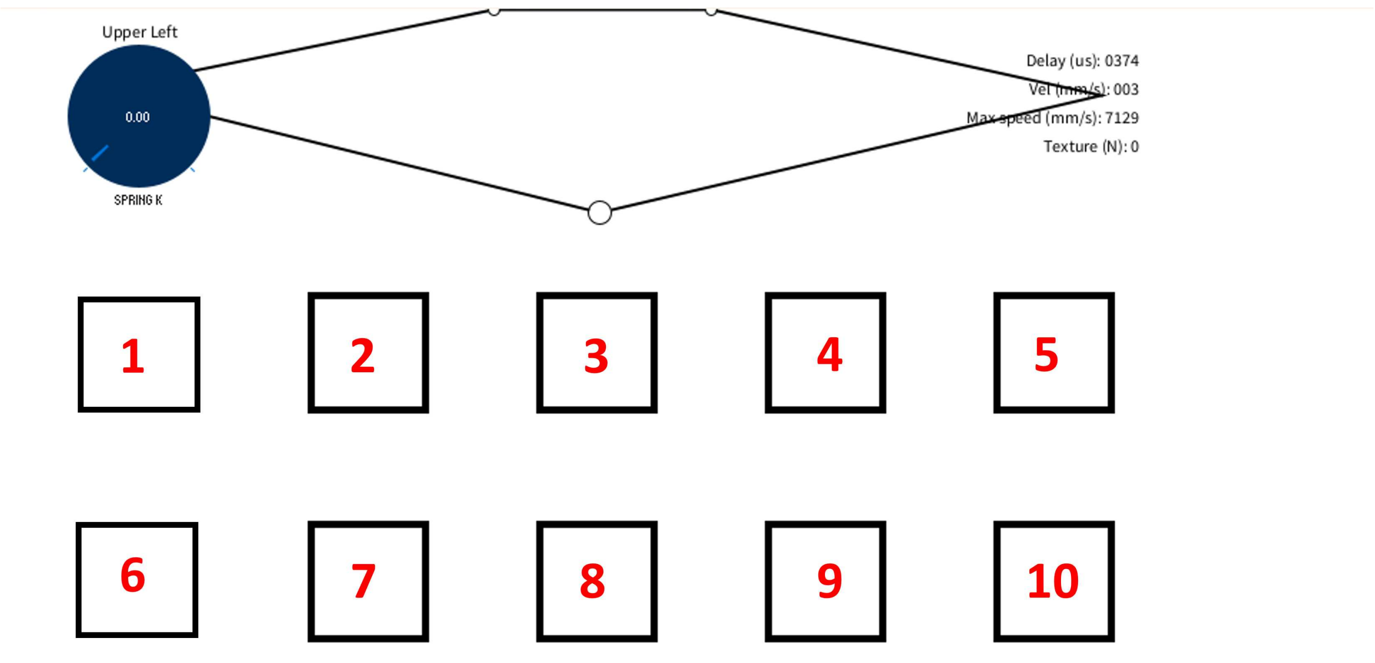


Fig. 5.2: The 2DIY interpretation of a line in the physical world translated to the virtual world.

The most current iteration of the 2DIY is not without its issues, and as a result, the project has pivoted to a qualitative approach as described in Section 3.3.4. Throughout the study, it was noticed that the encoded location of the end-effector seems to vibrate in place, even though physically it is not moving. I strongly suspect this is a result of the interaction of the encoder chip with the device. The encoder chip, which ultimately determines the location of the end-effector, reads the rotation of a magnet that is located within one of the gears from the latest hardware updates outlined above. Since the placement of the chip is not designed for this new iteration of the device, I believe that this would cause the chip to have a small level of uncertainty when determining the location of the end-effector, causing it to vibrate.

Throughout the study's execution, the pantograph arms were also replaced as a result of wear and tear. Traveling with the device, combined with constant assembling and disassembling, caused one of the pantograph arms to barely remain attached to the motor. Any significant force applied would cause the arm to slip in its socket, effectively breaking the intended force-feedback.

Haply Robotics has been instrumental in providing various fixes to their device as we test them. A design recommendation that I would make is to adjust the magnet location for the encoder chip as well as fix the chip's position to ensure constant and reliable readings. Barring a complete change in the 2DIY's form factor, a second suggestion regarding the frailty of the pantograph arms during travel is to simply provide a travel box that would protect the arms from further damage.



Delay (us): 0374
Vel (mm/s): 003
Max speed (mm/s): 7129
Texture (N): 0

Fig. 5.3: 2DIY testing suite. Each square will cause a ping sound whenever the end-effector enters and leaves as well as providing texture while within the objects.

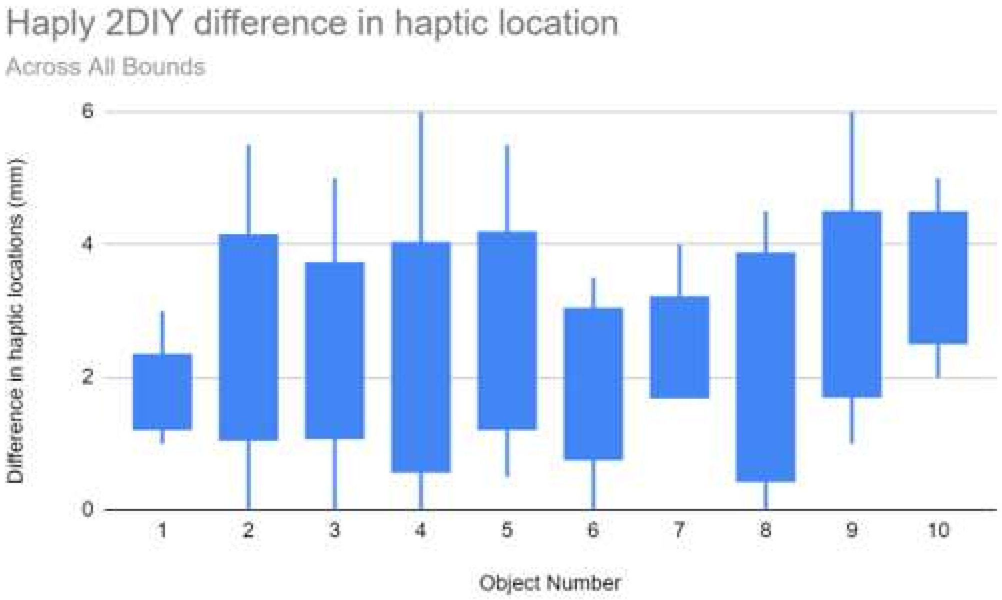


Fig. 5.4: Average offset across all edges for each haptic object arrayed in the virtual space. The center point of each candle is the mean offset, the top and bottom of the candle are the means +/- STD, and the top and bottom wicks represent the maximum and minimum offset values measured at that object across all devices.

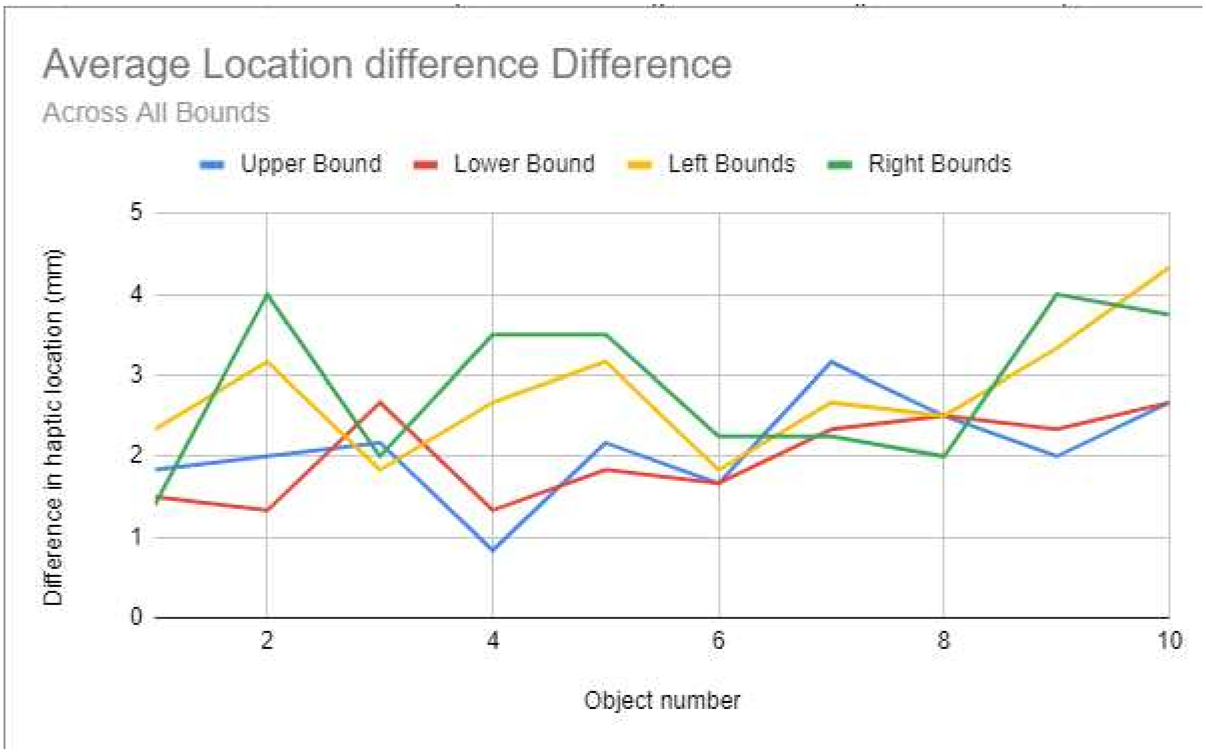


Fig. 5.5: Average Offset

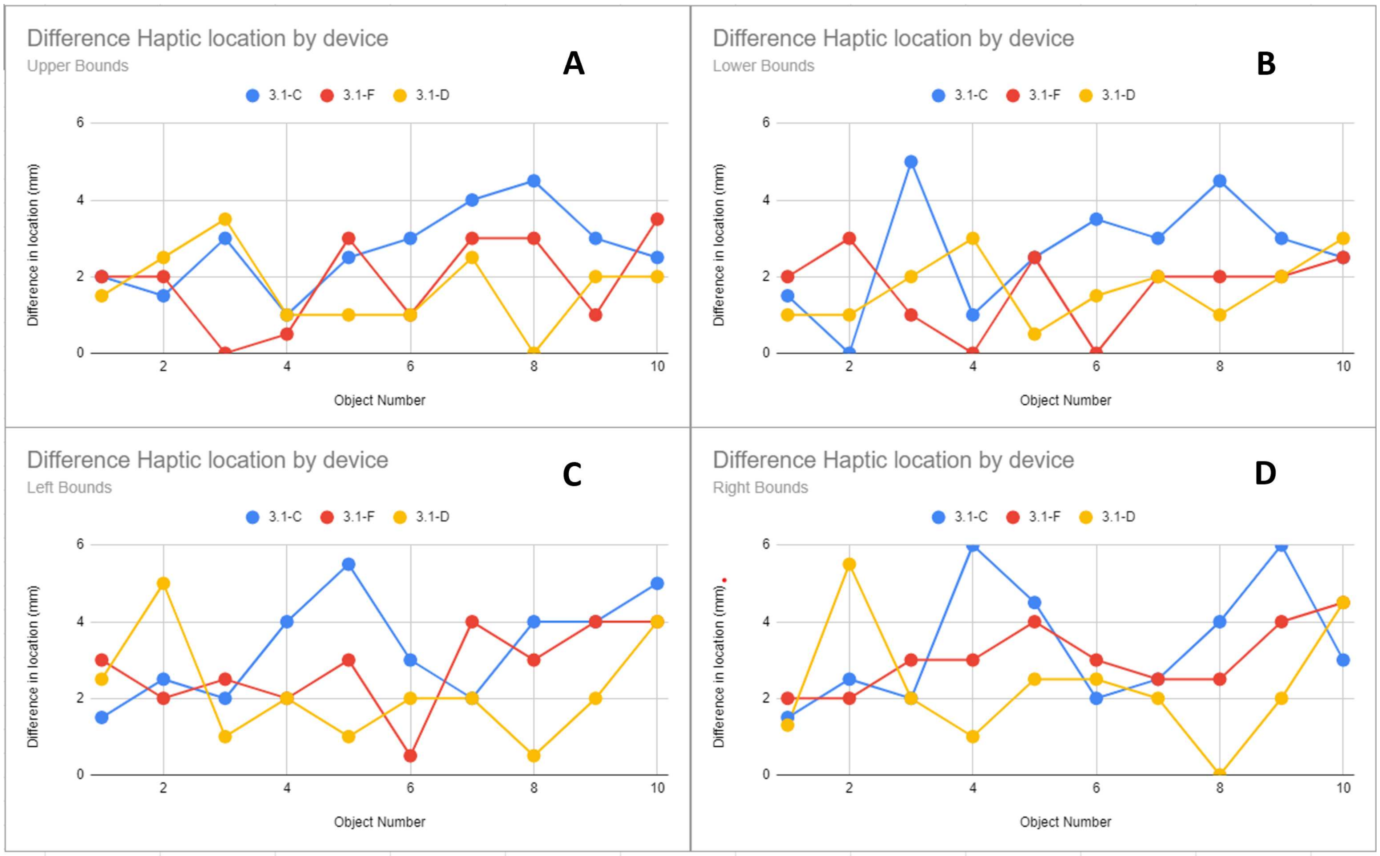


Fig. 5.6: Measurements of object edge location difference when approaching from different directions, each colour represents a different 2DIY. A, B, C, and D outline the difference in position of the top, bottom, left side, and right side of the haptic objects, respectively.

Questionnaires

Pretest Questionnaire

- How old are you?
- What is your gender?
- Are you left handed, right handed or ambidextrous?
- Do you have full range of motion in both the left and right arms and hands?

- If not, which arm or hand have a limited range of motion?
- Describe your degree of vision
- Are you congenitally blind or did you lose vision later in life?
- What age did you lose your vision?
- How many times a week do you use a computer?
- When presented with a photograph, what tools or methods do you use to understand its contents?
- How familiar are you with force-feedback haptic devices (in other words, devices that either guide your hand or provide resistance to eg. a force feedback mouse)? (5-pt Likert scale "Not at all to "Extremely")

Post Scene Questionnaire

- To the best of your ability and with as much detail as you can, please describe the scene.
- How confident are you with your description? (Likert 1- 5)
- If you were to guess, what are they(objects) doing?
- If you were to guess, where is this picture taken?
- Name one thing you enjoyed about this photograph? Why?
- Name one thing you disliked about this photograph? Why?
- What else did you want to know about the photograph?

After answering these questions, the experimenter "reveals" the scene by describing what is actually in the scene. At which point they ask the following:

- Name one thing that didn't match your expectations? Why?

5.0.0.1 Post Perspective and Post Experiment Questionnaire

- In the last two photographs, name one thing you enjoyed? Why?
- In the last two photographs, name one thing you didn't enjoy? Why?
- Was there anything about the photographs that made you frustrated? If so, what and why?
- Was there anything else you wanted to add about the last two scenes?

At the end of the experiment, the following questions are asked:

- If you had to pick one perspective, which one would you pick and why? Why?
- Do you have any other thoughts about this system or this study?