Measuring the Detection of Objects under Simulated Visual Impairment in 3D Rendered Scenes

A DISSERTATION SUBMITTED TO THE FACULTY OF UNIVERSITY OF MINNESOTA
BY

Brent Carpenter

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Adviser:

Daniel Kersten

September, 2018

Acknowledgements

Thanks to the Devil for pain and wisdom, and thanks to Humanity for inspiration

Dedication

This dissertation is dedicated to all those cyberpunks out there who fight against injustice and corruption every day of their lives. All controversy begins with doubt.

Abstract

A space is visually accessible when a person can use their vision to travel through the space and to pursue activities intended to be performed with vision within that space. Previous work has addressed the difficulty of evaluating the detection of objects in real spaces by observers with simulated visual impairments. This current research addresses the viability of using physically realistic 3D renderings of public spaces under artificially induced blur in place of the more resource intensive testing in the real spaces themselves while participants wear blurring goggles. In addition, this research illustrates the efficacy of a model that predicts portions of public scenes that an observer with simulated vision impairment would presumably fail to detect by comparing the predictions of missed space geometry to actual geometry detection failures by observers with simulated impairments. Lastly, this work also addresses how well simulated Low Vision observers can categorize the contents of scenes. Observer categorization rate is compared to several image metrics and the results indicate that average classification rate across Low Vision simulations can be predicted very well by knowing the averages of several different image metrics within each of the acuity blocks.

Chapter 1 of this dissertation is a literature review necessary for understanding the background of the state of the art of this research and an overview of the research itself.

In Chapter 2, an experiment is described in which object visibility was tested in a virtual environment with the goal of validating the use of 3D renderings as substitutive stimuli via comparing performance between the real and digital version of the same task (Bochsler et al., 2013). The objects were ramps, steps, and flat surfaces. Participants were normally sighted young adults who viewed either blurred or unblurred images. Images were blurred using a Gaussian filter on a Sloan Chart calibrated for the viewing distance of the experiment. Patterns of object identifications and confusions between the digital and physical versions of the task were highly similar. It is very likely that 3D renderings of public spaces when used in psychophysical tasks are effective substitutive stimuli for real spaces in object detection tasks. Avenues for parametric manipulations that might strengthen the argument are also explored.

Chapter 3 extends the use of physics based 3D renderings to simulations of visual impairment (Thompson et al, 2017; https://github.com/visual-accessibility/deva-filter). A model of visual impairment was used to simulate 3D renderings of public spaces under increasing levels of impairment. Participants were then asked to draw the edges and contours of objects in these simulations under several separate task conditions: draw the edges of doors, stairs, obstacles, or floor-wall-ceiling connections. As simulations of visual impairment deepened, observers struggled to find the correct object contours in each of the tasks. Also, as the simulated impairments deepened, observer data often more closely matched the predictive model: a system that puts a premium on sudden changes in luminance contrast. In the absence of context and meaning, simulated Low Vision observers tend to make false positive geometrical edge identifications when a scene has non- accidental incidences of strong luminance contrast edges such as bars of light and

shadows. The predictive power and utility of the model for simulating visual impairment is also discussed.

Chapter 4 contains a pilot experiment which seeks to understand how well simulated Low Vision observers can classify the category of blurry scenes shown to them. Observers were asked to perform a three alternative forced choice task where they must identify if an image is one of three scenes, and observers' classification accuracy was tracked across acuity level simulations. Several image metrics were calculated and regressed against classification accuracy of either single scenes or classification accuracy per acuity block. It was found that average classification accuracy within an acuity block could be predicted by knowing any one of several average image metrics of scenes within blocks and when regressed across acuity levels.

Table of contents

List of Figures

List of Tables

Chapter 1. Thesis Overview

"Thirty spokes unite around one hub to make a wheel. It is the presence of the empty space that gives the function of a vehicle. Clay is molded into a vessel. It is the empty space that gives the function of a vessel. Doors and windows are chiseled out to make a room. It is the empty space in the room that gives its function. Therefore, something substantial can be beneficial. While the emptiness of void is what can be utilized."

-Tao Te Ching, Chapter 11.

Imagine running an errand to a city hall in a major metropolitan area. The task requires you to cross busy streets, go up and down stairs, and squeeze your way through crowds of people. Each of these situations likely tests your patience more than it tests your ability to visually navigate them. However, if your eyesight is quite poor, even the simple act of crossing the street becomes a harrowing proposition. The uniform gray concrete stairs that lead to city hall represent a serious hazard to an individual with very poor visual acuity; this observer would have great difficulty finding a staircase inside what they perceive to be a large gray blob.

In this thesis, I focus on the problems of assessing the visibility and accessibility of public spaces for people with Low Vision. Ambulation through public spaces is often trivially solved for most individuals; however, visual navigation can be particularly difficult for a subset of the general population with poor eyesight. Low Vision is any eye condition, not correctable by glasses or contacts, that results in visual impairment. It is typically marked by exceptionally low visual acuity, contrast sensitivity, or a very small field of view (Leat, Legge, and Bullimore, 1999). Indeed, the low visual ability of Low Vision populations makes them particularly vulnerable to falls, stumbles, and collisions from failing to detect some navigation hazard (Lord and Dayhew, 2001). It is estimated that nearly 3 million Americans over the age of 40 currently have impaired vision. By the year 2050, it is estimated this number will more than double to nearly 9 million Low Vision individuals. (National Eye Institute, 2010). It is therefore important to address the emergent needs of an expanding special needs population and to assist in the creation of public spaces that facilitates navigation and mitigates falls for those individuals with Low Vision.

.

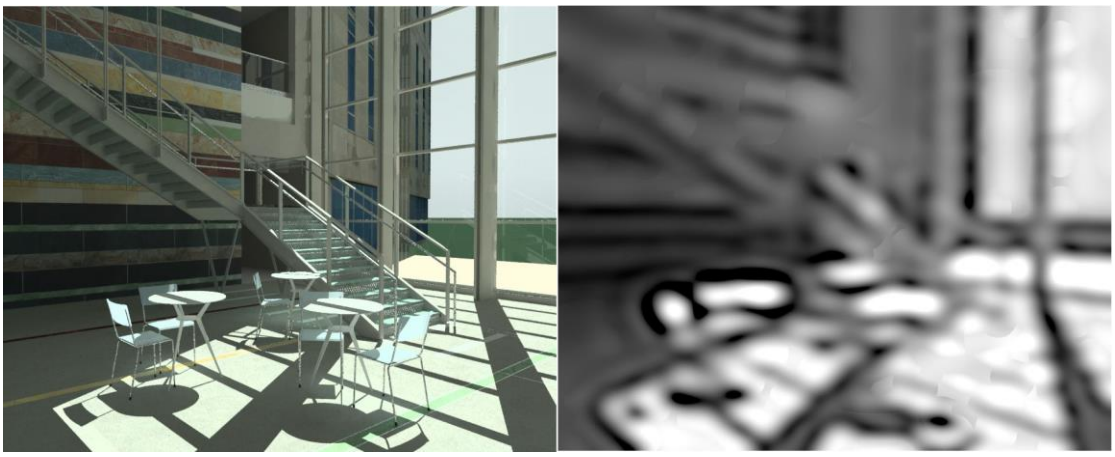


Figure 1. Two images of the same scene are shown side by side. The image on the right is a simulation of a public scene as seen by a hypothetical Low Vision observer with profound vision loss. The image on the left side is the same scene with no blur. Notice how the tables in the foreground blend into the stairs in the blurry image. In the blurry image, the landing where the stairs meet the ground is also difficult to place because of the chairs obscuring it.

As Low Vision populations expand, it is important to turn an eye to the public urgency of the Low Vision navigation problem; Low Vision individuals are at risk for falls in part because their navigation through public spaces is partially reliant on their vision. To be clear, individuals with impaired vision typically use navigational aids such as canes or seeing eye dogs for mobility or object detection, but, just as often, those with Low Vision use their residual eyesight to detect imminent hazards (Goodrich & Ludt, 2002). Of course, obstacles that impede ambulation could appear in indoor or outdoor spaces, but the structure of the public domain itself is much more easily modifiable than the structure of the private one; a concrete staircase in a public area benefits from the gaze of the conscientious public in that the utility of access to all citizens is readily apparent. In the general population, most falls occur in outdoor rather than indoor areas despite individuals spending far less time outside than inside (Bergland, Jarnlo, and Laake, 2003). Interviews with older and middle-aged adults, an age group most prone to falls, concerning their most recent injury due to a fall often reveals that these individuals fell while walking on uneven, slippery or poorly maintained public walkways (Li et al., 2006). Similarly, individuals with Low Vision, such as a poor field of view, have a significantly decreased ability to navigate around pedestrian obstacles, and, as such, they are particularly at risk for falls and injuries (Hassan, Lovie-Kitchin, and Woods, 2002). A loss of contrast sensitivity and difficulty perceiving depth cues are often considered primary risk factors in falls for Low Vision individuals (Lord, 2006). However, statistical analyses of likelihood of falls and visual disability tend to find little connection between low visual ability and likelihood of falling but the same analyses do find a connection between Low Vision physical inactivity and increased likelihood of falling in public (Lamoureux, 2010). In other words, the most likely person to fall in public in a Low Vision population is the one whom receives little exercise, has very poor contrast

sensitivity, and tends to be in poor physical condition: the person most likely to be seriously injured by a public fall. It is therefore a public responsibility to do the just dues that would reduce the likelihood of falls due to some poorly maintained or designed public infrastructure.

Many public falls are likely preventable had there been more care taken to highlight sudden elevation changes, simulation of lighting to find areas where bright lights mask changes in ground continuity, placement of footpath delimiters that show the difference between walkable paths and decorative constructions, and so on and so forth. To that end, designers of civic structures have clear instructions for the creation of safe and accessible structures: highlighting sudden drops in height, rules for the average step height on stairs, and so on (OSHA, 2017). The goal of such endeavors is to foster a public space that minimizes the likelihood of injury while creating a space amenable to the activities of day to day life. However, when action is taken to prevent specific individuals from falling, most fall prevention interventions designed to reduce falls do not focus on changes or modifications to public environments. Instead of modifying the public domain into accessible spaces, fall prevention is often the purview of families and personal physicians (Gillespie et al., 2003). The emphasis on personal solutions to private problems creates the appearance that Low Vision needs are being met by conscientious efforts in the public and private domains. However, even with considerable effort taken by civil engineers, architects, and personal physicians to create safe and walkable public spaces that prevent falls, many Low Vision individuals are isolated in their homes because of their fear of falling and hurting themselves in public (Willis et al., 2012).

Isolation due to poor visual ability is a solvable problem if the public environment is modifiable in such a way that one can both predict and alter areas of public spaces to prevent falls and increase salience of pertinent obstacles. Low Vision reduces mobility, in part, due to a person's belief that public spaces are not visually accessible; individuals that can only correctly identify small portions of a space at a time often adopt extremely cautious approaches to navigating that space (Timmis and Pardhan, 2012). Improving the mobility of Low Vision populations requires that one adopt a strategy of prediction: what spaces, obstacles, constructions, lighting fixtures and so on produce the greatest visual confusion? However, Low Vision is a heterogenous visual condition that can arise from many different causes, and, as such, the visual abilities of any one Low Vision observer could vary greatly from even their aged matched peers (Leat et al., 1999). The task of improving Low Vision mobility through public intervention then becomes twofold; one must know the capabilities of an observer that can take multiple forms and then one must predict the mistakes that this multifaceted observer might make in a public space that itself could take any form. To that end, the works proposed here seek to lay a foundation for the integration of both psychophysics and image processing techniques that together test the veracity and feasibility of predicting Low Vision behavior in a wide variety of public areas. However, there's the rub; how can one predict the undetected obstacles of individuals with limited visual abilities?

It might seem straightforward to predict the parts of public spaces that would provide pause during Low Vision navigation. One might simulate Low Vision perception of a public space via image processing techniques such as blur, contrast adjustment, or placement of apertures: low acuity, contrast insensitivity, and poor field of view

respectively. However, civil engineers and architects are intimately familiar with their architectural designs, and familiarity with a scene or object increases the likelihood of making incorrect object recognitions when that same scene is seen under blur (James et al., 2000). The matter of predicting hazardous parts of a public space is complicated by the variability of visual ability in Low Vision observers; Low Vision can result from any combination in almost any level of severity of acuity loss, contrast sensitivity loss, or visual field loss (Cheung and Legge, 2005). To make matters more complicated, improving Low Vision object detection is not as simple as painting a yellow strip across the surface of some object. Instead, there is an interaction effect between the lighting in a space and the relative contrast of an object with its background such that some lighting arrangements might influence Low Vision object detection even when the object has strong contrast with its immediate background (Kallie, Legge, Yu, 2012). A single public scene can be viewed from many angles, and this produces a need to step inside of the scene and view it with the eyes of a Low Vision observer because some objects might be detectable from one view and not another. For example, a downwards staircase might be almost invisible while approaching it, but the same staircase could be particularly easy to spot when seeing it from the side. Predicting the likelihood that a Low Vision observer will see a hazard in a public space is nontrivial given the great heterogeneity in observers and spaces to be navigated.

Predicting the hazardous areas of a space requires that one somehow steps into and views that space with the eye of a Low Vision observer. A 3D space can be seen from many viewpoints, and these viewpoints can be captured as snapshots of that scene: as images. It would be prudent then to manipulate images of a public scene such that the manipulation represents the ability of a Low Vision observer with a prescribed visual loss of acuity, contrast or visual field. We focus here on the first two, quantifying an observer's sensitivity to spatial detail (acuity) and contrast.

The visual sensitivity of an observer can be characterized by their ability to just detect sinusoidal patterns that vary in spatial frequency and contrast amplitude (Robson, 1966; DeValois and DeValois, 1988). Visual sensitivity can be characterized by the contrast sensitivity function (CSF); the reciprocal of the minimum visible contrast of a static sine wave grating as a function of spatial frequency. The path would then appear clear in that the way forward would be to characterize a general loss of spatial vision, acuity or contrast sensitivity loss, with a general CSF model, and to then use that general model of vision loss to alter images and simulate Low Vision It is currently unknown if manipulating images with such a model would adequately capture Low Vision object detection patterns. Observe the images in figure 1. The images are manipulated according to such a model, but how close of an analogue to Low Vision such a manipulation is remains an open question

When considering the future of Low Vision research, it will also be necessary to identify how Low Vision individuals interact with objects as well as how they identify them. This is because knowing the identity of an object does not mean that one knows what to do with it, and, likewise, knowing the purpose of a hammer does not mean that one will use it well (Goodale & Milner, 1992). An example of this can be seen in patients that are clinically blind whom still successfully grab objects that they don't consciously see: a phenomenon known as blindsight (Perenin and Rossetti, 1996). As a pertinent

example, if a Low Vision person successfully finds a staircase while walking through town, it is still possible for a fall to occur due to some misstep on that staircase. When an observer locates a staircase, they must then figure out how far to raise or lower their foot, and a miscalculation during the motor movement could result in a serious error. Low Vision research must necessarily include an investigation into the nature of affordances in pedestrian obstacles: the traits of that object that facilitate the valence of its own usage (Greeno, 1994). Before identifying an object's affordances for usage, an object must first be located, and even the intuitively simple act of locating an object is complicated by a priori knowledge from the observer. For example, the perceived depth of an object can change as a function of its location relative to a horizon. Blurry objects sitting above a visual horizon are estimated to be farther than their actual position (Rand et al., 2011, Rand et al., 2012), and this introduces the possibility that Low Vision observers might bump their heads on low hanging obstacles. If such a low hanging object had a design that facilitated its own recognition and navigational interaction, a Low Vision observer would be less likely to bump into it as the navigational paths around or under it could be identified with ease and without anxiety.

The future goals of Low Vision research should therefore be two pronged: how are Low Vision observers locating pedestrian objects and how are these observers able to safely interact with these pedestrian obstacles? Future research would do well to focus on those qualities of pedestrian obstacles that both enhance detection and reduce the likelihood of pedestrian ambulation errors: likely a result of enhancing the affordances pertinent to that pedestrian object. The work here aims to build a pathway towards simulating pedestrian navigation and usage of public spaces by first exploring and assessing the visual accessibility of simulated public spaces.

In the following chapters, I describe my research on three projects that were conducted as part of an interdisciplinary multi university project known as "Designing Visually Accessible Spaces" (DEVA). A primary goal of this project is the development of tools, methods, and techniques that aid civil engineers, architects, and interior designers in the creation of public spaces that are accessible for people with impaired visual abilities. The ultimate goal of which is the proliferation of design principles that allow public spaces to be utilized by all no matter how complex the design of the space or the severity of impaired vision in the observer. For the Low Vision observer, this means that the key features of a space such as doors, stairs, or obstacles in the walking path are easily seen. The DEVA project has already designed tools and techniques for manipulating images and simulating low visual abilities, but it is unknown if these simulations are accurate predictors of human performance. Computerized tools that the DEVA project has developed to use simulations of low visual ability that are accurate to real low visual ability (Thompson et al, 2017; https://github.com/visual-accessibility/deva-filter), and these simulations of Low Vision are used to predict the detection of the geometry of public spaces such as stair edges, door frames, walls, and so on.

The current scientific goals involved in DEVA's desire to aid Low Vision mobility requires a combination of psychophysics and image processing techniques that allows one to step into a scene, produce a Low Vision simulation of that scene, and predict what is and is not visible to that observer. The manipulation and analysis of

digital images is a productive avenue for predicting objects that are hazardous to Low Vision observers, but this is only the case if there is some certainty that image processing techniques and Low Vision simulation produce behavioral patterns like real Low Vision behavior. Once it is known that image processing suffices for accurate simulation of Low Vision behavioral tendencies, it will be necessary to quantify the extent to which simple processing techniques are predictive of Low Vision object detection and to understand if the techniques of DEVA's quantification of visibility outperforms any predictions one might build by deriving image statistics in a scene. The final step in bridging the gap between DEVA's predictions of Low Vision visibility and the content of an image of a public scene that might be driving observer behaviors is to understand the extent to which observers must sample the image space before making accurate hazard identifications.

To those ends, three separate experiments were conducted that each aimed to understand the different aspects of Low Vision obstacle detection or identification while also attempting to meld psychology with computer science techniques. The problem of predicting the unseen features of a scene is broken down into several experimental problems that, hopefully, spark the foundation for future research into this endeavor. The scientific issues are threefold, addressed in chapters 2, 3, and 4.

Chapter 2, addresses the question if one can sufficiently emulate Low Vision behavior with both simulations of real spaces and simple image processing techniques. Low Vision observers can be difficult to pull into the lab as participants, and the number of public spaces and potential variations in the designs of these spaces that confound Low Vision navigation is nearly limitless. We believed that data from previous Low Vision based experiments could be reproduced in the lab by utilizing both simulations of Low Vision and 3D renderings of prior Low Vision stimuli. To that end, we conducted an experiment that replicated and reproduced a prior Low Vision experiment performed by Bochsler et al (2013), and we hypothesized that, if the experiment was reproduced exactly in terms of both visual angle and similar levels of reduced acuity, participants will produce similar patterns of object recognition under simulated low acuity as seen in the previous experiment. This was a simple yet crucial first step towards studying the visibility of yet to be constructed public spaces that could take almost any shape or form: with the validation of 3D renderings and tonemapping techniques as suitable for Low Vision psychophysics.

In chapter 3, an algorithmic image analysis pipeline developed by DEVA (Thompson et al, 2017) was used to analyze images of public scenes for points of geometry that might be visible or invisible to some simulated Low Vision observer. Those predictions of visibility were compared to visibility scores obtained by having participants trace geometry in the same scenes at the same blur. Participants were asked to trace the contours of obstacles, doors, steps, and floor wall connections, and it is assumed that every tracing represents the participant's inference, correct or not, about the presence of underlying scene geometry. The DEVA analysis process utilizes both a general purpose Low Vision simulation model and algorithmic image processing pipeline for assessing edge visibility that relies on quantifying the distance between underlying geometry edges and luminance changes diffused across the image space by blur. The nature of quantifying the distance between luminance boundaries and geometry edges and the accuracy of those quantification metrics for predicting human visibility are

explored. The DEVA filter's prediction was compared to globally derived image statistics such as the average contrast in a scene, and the goal was to determine if simple image statistics could be used to predict visibility more efficiently than metrics that produce visibility scores by scaling distance to tracing or distance to luminance boundaries. We concluded that the DEVA filter's technique of measuring distance between selections of obstacle contours and underlying geometry edge was predictive of geometry visibility, and we also found that the DEVA filter process outperforms the predictive power of simple image statistics.

From the tracing experiment, it appeared that observers were sampling the image space for relevant information to their tracing decisions. Observers would occasionally detect parts of scenes that were predicted to be invisible by the DEVA filter, and, most notably, those predicted to be invisible geometry pixels were often surrounded by predicted to be seen geometry pixels. This would often occur in door frames or in stair cases where some part of the object had portions that were too low in contrast to be clearly seen, yet participants managed to connect the dots and trace over the invisible geometry pixels by drawing lines from neighboring visible points. This behavior of "completing the form" of some nearly invisible portion of an object tended to occur in situations where the object was uniform in its shape such as a door or a long platform.

Chapter 4 explores how simple image metrics and derived geometry boundary visibility scores can predict the ability of observers to categorize the content of images. A pilot experiment is described in which observers were required to identify whether images presented to them contained either doors, chairs, or neither doors or chairs under three different Low Vision simulations: moderate, severe, and profound. The goal was to predict the classification accuracy of observers by deriving several image metrics from the images used in the experiment. High spatial frequency, root mean square contrast, and geometry visibility scores were derived for individual scenes, and linear analyses were performed on both classification performance for single images and average classification performance within acuity blocks to each of the three metrics across all acuity levels. We found that classification accuracy could be predicted by the average of some metric for images within an acuity block, but observer performance was not predictable just by knowing the derived metric for any one single scene. The implications for predicting Low Vision scene classification accuracy are discussed.

Altogether, these experiments demonstrate the utility of 3D renderings as viable stimuli substitutes for real public spaces, the utility of predictive modeling of visibility of edges in renderings of public spaces, and the quantitative effects of information loss on object visibility. The works here both qualitatively and quantitatively demonstrate the strong effect that visual impairment has on observers before they begin ambulating in that space. In addition, the building blocks for future research workflows are present here in these collective works: effective stimulus substitution in a high dimensional stimulus category, predictive modeling efficacy of visual impairments in a heterogenous population, and basic quantitative explanations of predicting simulated Low Vision observers' scene classification accuracy.

Chapter 2: Validating the Use of 3D Renderings in Low Vision Psychophysics

Background

Imagine running an errand to a local gas station. The act of walking into and out of the station is seemingly trivial for most individuals; however, even this simple act could become unusually difficult if one had exceptionally poor eyesight. Most gas stations themselves are positioned on a slightly higher concrete platform to delineate drivable space from pedestrian areas. The difference in elevation may or may not be highlighted for pedestrians. In either case, a pedestrian with very poor eyesight must somehow detect a sudden increase or decrease in height. In the latter case, the low sighted pedestrian must detect a concrete curb against a concrete background. It is a harrowing proposition when failure might result in personal injury.

To improve the visual accessibility of some future public space, it is necessary to know how and when Low Vision observers make object recognition mistakes. Information about navigational behavior and hazardous object detection is often gathered by having participants walk through an obstacle course. However, these obstacle courses are cumbersome to create and difficult to manage; in addition to this, it is not always clear what obstacles Low Vision observers fail to see. The construction of obstacle courses also poses a severe limit on the number of testable obstructions, but the design possibilities that architects have are almost endless. It is necessary to recreate real world navigation hazards in the laboratory without having to reconstruct every possible staircase, sidewalk, or hallway, and it is also critical to know if such a recreation results in stimuli that preserve the basic components of the scene that lead to patterns of visibility for Low Vision observers.

With three-dimensional (3D) rendering, it is possible to reconstruct any object or scene that one could imagine. It is also possible to use software such as Radiance rendering to create physics based renders of luminated 3D surfaces (Larson and Shakespeare, 2004). However, Radiance produces images in a high-dynamic-range (HDR): luminance values that far exceed the range of displayable RGB values on a standard display device (Devlin, 2002). To display HDR images on a low dynamic display, the image must be tonemapped; the luminance values of the HDR image must be converted into an RGB space within the displayable range of the monitor (Devlin, 2002; Drago et al., 2003).

The process of tonemapping often involves the loss of information at either extreme of the dynamic range of the image. This poses a clear problem for use of 3D renderings as a substitute for construction of obstacle courses in testing visual accessibility. If contrast sensitivity is a strong predictor of an observer's ability to detect an object, it is possible that the tonemapping and display of realistic 3D renderings could impair the ability of observers to detect subtle changes in luminance while under severe blur. The human eye can accommodate a lighting ratio of surfaces of 10,000 to 1 in a single scene, but a typical CRT monitor can only display simultaneous tones that vary in luminance in a ratio of 100 to 1 (Krantz, 2000; Ghoudrati, Morris, and Price, 2015). The eye outstrips the typical research monitor in representation of luminance by a factor of 1000 (Ferwerda, 1996). In order to improve contrast ratios on a monitor, it would be necessary for monitors to be much brighter, but most electronics consumers appear to be

satisfied with the current state and progression of monitor technology. This is likely because using a brighter monitor for better contrast ratios would produce a situation where users are staring into a lightbulb just to see more realistic presentations of real world scenes.

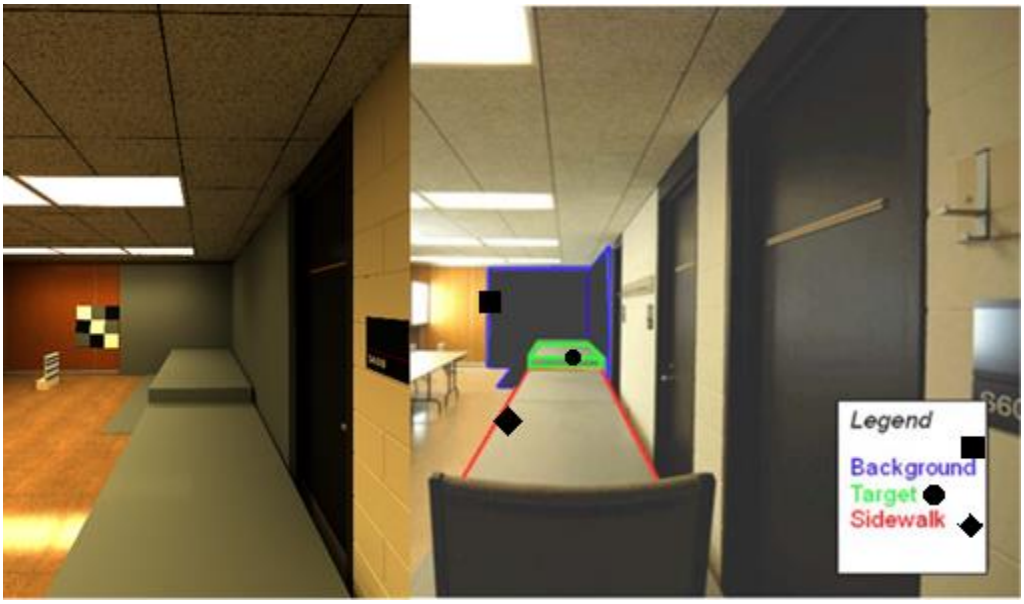


Figure 2. The experimental set up used in Bochsler et al. 2013 is shown on the right and the 3D rendering is shown on the left side. The scene on the left was rendered in Radiance 3D to be a physically accurate model of the stimuli and room. A photometer was used to measure and precisely model every surface of the experiment. The images depict a "Step Up" stimuli, and the components of the stimuli themselves: the target area, the sidewalk leading to the target, and the background around the target. Both experiments used 5 such stimuli: a step up, a step down, a ramp up, a ramp down, and a flat continuation of the sidewalk.

It is then critical to confirm that tonemapping and display of physics-based 3D modelling does not affect an observer's ability or, more importantly, their inability to detect objects while under simulated Low Vision. The present study investigates the ability of observers to identify common pedestrian hazards: ramps, steps, and flat surfaces. The current experiment is a 3D simulation of a previous experiment conducted on a real set of physical stimuli (Bochsler et al., 2013). A method for the comparison of real and physical detection data is explicated. The objective of the experiment is to validate the utility of 3D renderings as a medium for psychophysical experiments on visual accessibility, and to delineate a clear path for future researchers to use physics-based renders as a channel for psychophysical stimuli. The validation of 3D rendering as a stimuli medium is accomplished via the comparison of patterns of accuracies and errors between hand-built scenes and digital scenes; If tonemapping and display of HDR 3D images does not affect a simulated Low Vision observer's ability to detect navigation hazards, the patterns of accuracies or errors generated in the digital task should match the same patterns generated in the physical task.

Methods

Participants. 6 graduate students from the University of Minnesota took part in the experiment. The experiment was approved by the Institutional Review Board of the University of Minnesota. Each participant completed a single experimental session that lasted about one hour, and each participant completed all 6 acuity levels during their participation. The participants all had normal or corrected to normal vision and they were compensated monetarily for their participation.

Stimuli. The stimuli were three-dimensionally generated images of a room originally used in an obstacle recognition study in a previous Low Vision object recognition experiment (Bochsler et al., 2013). The images were generated in Radiance 3D Rendering (Larson and Shakespeare, 2004). The images show an elevated floor construction partitioned in one of five shapes: a step upwards, a step downwards, a ramp upwards, a ramp downwards, and a flat surface. The images were displayed with 800 x 1200 pixel resolution.

The stimuli were blurred with a Gaussian filter to match various acuity levels. There were six acuity levels ranging from logMAR 1.0 to 2.0 in steps of 0.2 LogMAR. The Gaussian filter was calibrated to blur images to these logMAR values by first generating a Sloan Acuity chart calibrated for the experiment's viewing distance (e.g. each line corresponded to the appropriate acuity for the participants' distance from the computer monitor) (Bailey and Lovie, 1976; Field and Brady, 1997). The chart was subjectively blurred until a corresponding logMAR's line could not be read and the line before it could only barely be read. The parameters for that Gaussian blur were used to blur the rendered images only after those images had been resolution matched to the Sloan blur chart. Failure to resolution match the renderings to the chart would produce differential blur simply due to the mechanics of a Gaussian filtering process (Olberholzer et al., 1996). A Gaussian blur kernel attenuates image information over a rolling window within that image that cascades across the image surface; a blur kernel of some size for some image will not have the same blurring effect for the same image at a different resolution.

Presentation of stimuli. During the task, stimuli were presented on a 20-inch CRT monitor (IBM flat screen, resolution 1600x1200 pixels, refresh rate 120 Hz). Participants were seated 13 cms from the screen so that their field of view of the stimuli would encompass the same retinal space as would have occurred in the physical task. The images spanned a 75 by 50 visual degree field of view. The monitor was gamma calibrated with a photometer. The output of luminance values from the monitor were linear.

Radiance produces high-dynamic-range (HDR) images. HDR images contain luminance values that far exceed even the brightest pixel value that a monitor can display. This poses a serious technical problem as it is desirable to replicate the exact luminance values of real world scenes, but a standard LCD research monitor can only

display luminance values up to about 300 cd/m^2 or as low as 5 cd/m^2. However, most monitors cannot display very bright or dark patches simultaneously, yet the human eye is very easily able to accommodate such luminance ranges.

The problem is solved by tonemapping; transforming the image data from luminance values to RGB values and compressing some of the information such that little behaviorally salient information is lost. Most of the dynamic range in an HDR image comes from a lightbulb, sun, or bright specular highlight. These parts of images can exceed 1000 cd/m^2, and a tonemapping algorithm will cap the displayable values in RGB space such that all the pixels corresponding to the lit surface will be pure white; typically, little information of behavioral relevance is lost. For this experiment, stimuli were tonemapped via Radiance 3D rendering (Ward & Shakespeare, 1997). The tonemapping was linear up until luminance values were twelve times the median luminance for the scene, after which the values were capped a maximum luminance value.

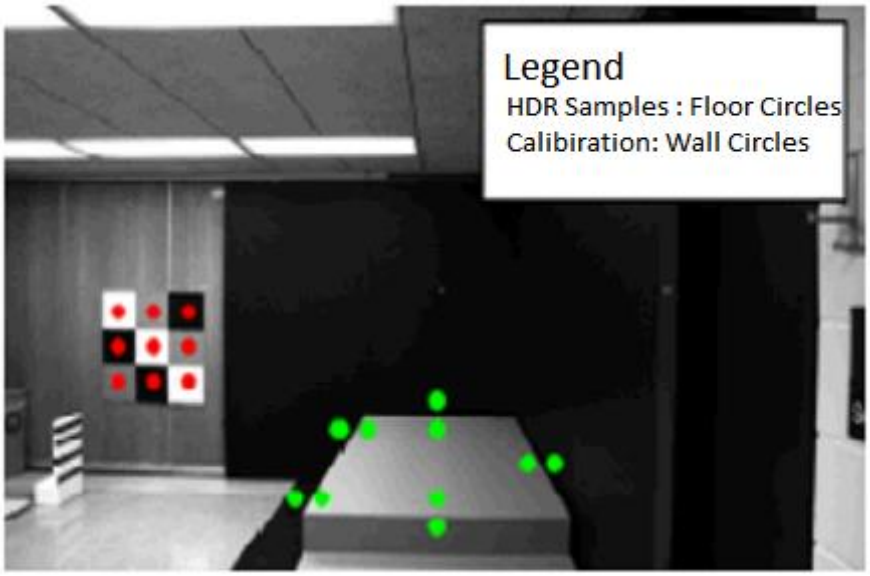


Figure 3. The luminance measuring points read by a photometer both in the physical room and in the digital simulation. The green points represent the positions of luminance pairs used to calculate contrast points of scene stimuli. The red points represent a calibration pattern used to calibrate the photometer. Green points were paired together to produce contrasts at the corners of each object. The image here shows the contrast points gathered for a step-up stimulus, but the process was repeated for all 5 of the stimuli. The contrast points both in the physical and digital experiments were paired both by location and by stimulus type for the linear analysis.

To compare the display of tonemapped HDR images to the original scene, a linear regression analysis was performed. Contrast ratios were measured at 10 points for each of the five classes of stimuli: step up, step down, ramp up, ramp down, and flat surfaces. The ten points measured for luminance values were points on and off the perimeters of

the various stimuli. This was done to acquire contrast values for the stimuli themselves against the background. Similar data were available for the stimuli in the physical room, and a linear regression was performed between ordered pairs of the contrast ratios for the 5 stimuli classes. The 50 luminance values produced 25 contrast ratios which, in turn, produced 25 ordered pairs of contrast values between the physical and digital displays. Ideally, there would be a 1-to-1 correspondence between the contrast ratios in the physical and digital scenes, and this would be evident if the regression line passed through 0: no average bias in contrast ratios towards either scene. The intercept of the regression line was found to be an insignificant parameter for the model: $t(1,23)=0.632$, $p= 0.534$.

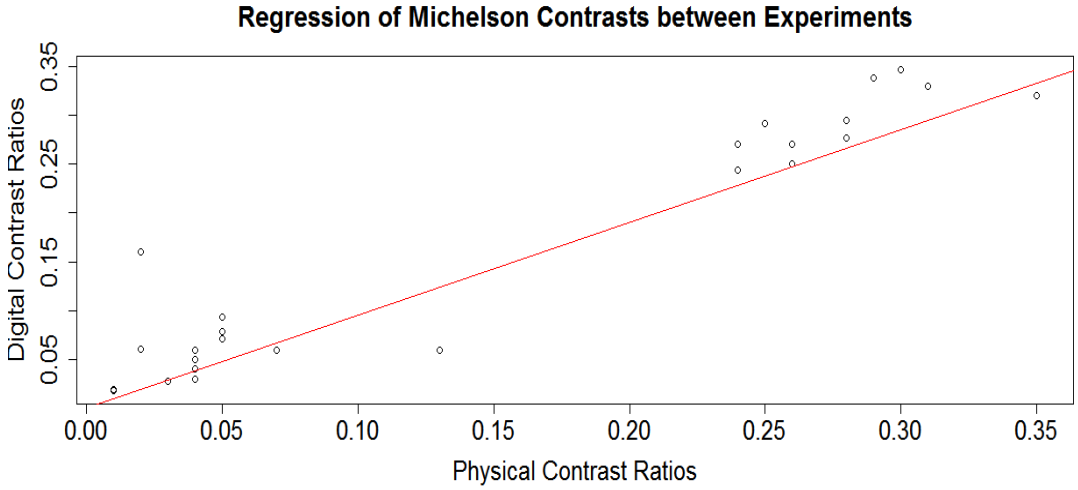


Figure 4. The linear analysis of paired location contrast points gathered from photometer measurements in both digital and physical scenes. The data shows strong clustering and high agreement within both clusters. The linear analysis shows little disagreement between the location specific contrast ratios of either digital or physical stimuli on a scene by scene basis. Most notably, both scenes have the same range: 0.35. The digital scene is presented on a low dynamic range monitor; the location specific contrasts match between the physical and digital scenes despite a change in dynamic range between the two.

It is also critical to emulate the exact field of view of observers in the previous experiment; such that, the observers in the digital version experience an identical subtension of the stimuli in their retinal image. To do this, the viewing angle participants had in Bochsler et al was recreated as close as possible in the digital simulation: both the field of view of the scene and the visual angles occupied by the stimuli. The participants in Bochsler et al stood on a stimuli platform and viewed the ramps and steps stimuli from 10 feet away, and these participants viewed these obstacles through welder's goggles that slightly restricted their vertical and horizontal field of view. By wearing these goggles, standing in the same spot as previous participants and looking at the stimuli, one can

measure the horizontal and vertical extent that participants had while wearing the welding goggles. With the welding goggles on, participants had a 77-degree visual angle horizon and a 100.8-degree visual meridian. Therefore, the digital simulation displayed the stimuli images such that participants would have these images would occupy the same field of view: by sitting at 13 cm away from the monitor. If the display of images in the digital simulation is true to the physical form, then certain ratios of visual angles should be maintained in the digital image that exist in the physical room. For example, in the physical step up, the top edge of the step occupies 22.6 degrees of visual angle and then the stimulus platform continues until it contacts a wall and this edge occupies 12.68 degrees visual angle. Both the step edge and the portion where the stimulus meets the wall are the same size, and, therefore the ratio of the two shows the degree to which the visual angle of the stimuli shrinks in perspective projection: 0.5611. In the digital scene for the same potions the visual angles are 12.8 and 22.9 respectively and the ratio of the two is 0.559. The ratio of the ratios therefore shows the degree to which this foreshortening has been maintained in the digital display relative to the physical room on which it is based and this ratio is 0.9963; the digital display is very nearly a 1 to 1 facsimile of the physical room in terms of viewing angle.

Acuities simulated in the current experiment ranged from LogMAR 1.0 to 2.0 in steps of 0.2. Simulated low vision acuities in the Bochsler et al (2010) experiment were 0.8293 and 1.6532 LogMAR. The new range of simulated acuity loss encompasses the previous experiment's range and extends it slightly to much more severe acuity loss. This was done because the Bochsler et al experiment tested the visibility of ramps and steps on both Low Vision and simulated Low Vision groups, and the actual low sighted participants had acuities that ranged from 0.98 to 2.18 LogMAR

Experimental Procedure. Prior to running in the experiment, subjects were shown the original unblurred stimuli, and the subjects were run on a short 25 trial guided training session. Each of the 5 stimuli were presented 5 times without blur. The training session was performed to acquaint the subjects with the slight differences between the images, to ensure perfect identification with no blur, and to train subjects on the experiment's interface. Data was collected via a mouse interface. After a stimulus trial, subjects had to click a box to indicate their possible answers: step up, step down, ramp up, ramp down, or a flat surface.

Each of the 5 stimuli were presented 15 times for each acuity level and, ranging from logMAR 1.0 to 2.0 in steps of 0.2, there were 6 acuity levels. For each participant, there were 90 total stimuli presentations. Participants were run through all acuity levels in a counterbalanced Latin square design. Subjects completed each acuity level in discrete blocks such that there was no overlap in stimuli presentation between acuity levels.

Within an acuity block, blurred stimuli were presented in random order. Each stimulus was onscreen for 4 seconds and subjects were not instructed to fixate anywhere on the screen. After scanning the stimuli for 4 seconds, subjects were given a visual prompt that asked them to click a box next to the name of the stimuli they just saw. The process of random presentation repeated until all 5 stimuli had been presented 18 times.

The acuity block then ends and the next acuity block begins until a subject has completed all acuity blocks.

Analysis Procedure. The experimental procedure generates a data structure known as a confusion matrix; this is a matrix created for classification data where correct answers are binned in diagonal elements. The off-diagonal elements contain the confusions that the classifier had when performing the task. Rows correspond to the object presented and columns to the responses given. Therefore, the [1,2] element of the matrix would be the number of times the classifier was presented with the first object and responded with the second. However, the [2,1] element would be the opposite: the number of times the classifier was presented with the second object and responded with the first.

In machine learning, a confusion matrix is often used to evaluate classification algorithms. For these applications, the confusion matrix is usually a 2x2 matrix where the columns represent the predicted yes or no answers and the horizontals represent the ground truth of whether something really was or was not present. The simple 2x2 matrix allows one to calculate many different values about the classifiers behavior such as the accuracy, specificity, precision, true positive rate, and so on.

The data from this experiment is compiled into a 5x5 confusion matrix. There are 5 objects to identify, and participants can respond with any of these 5 objects during the answer phase of the experiment. The larger than normal matrix creates some interesting complications. For example, it is not actually possible to create an ROC curve when working with a confusion matrix that has a rank greater than 2. The reason is that the notion of a false positive or false negative only has meaning when there is a method of juxtaposition against a true answer. If a participant is presented with object 1 and responds with object 3, does object 3 represent a false positive or a false negative? The same could be said for any other combination in the matrix for this experiment. The notion of false positive and negative has no meaning in this paradigm and, thus, it is difficult to evaluate the general accuracy of observers in the task let alone the similarity of observers between tasks.

The confusion matrices in the current experiment are generated by participant response data in an object recognition task. Participants must identify hazards on a path, and these hazards could be one of 5 things: a step up, step down, ramp up, ramp down, or a flat surface. When a participant is shown an object, they must respond with one of these 5 identifiers, and that response is counted and binned in the appropriate element of the confusion matrix.

Each participant generates six confusion matrices: one for each level of visual acuity tested. When a participant finishes the experiment, their confusion matrices are summed together. The summed matrices represent the average classification performance across all acuities for that participant. There are practical reasons to sum across acuity levels. Performance on any one level of acuity has a large variance: some participants are good at the task and some are below average at it. Averaging across acuities gives a

measure of their overall performance. In addition to this, Low Vision is difficult to model, and people have Low Vision for a multitude of reasons. Any sampling of Low Vision behavior in some psychological task is either highly individualized or averaged across a wide range of visual acuities.

To have a sense of the overall average performance across participants, a master confusion matrix was generated from all participant data. The summed values of each participants' data are, themselves, added together. Each participant generates 6 confusion matrices: one for each acuity block. All the confusion matrices are summed together and averaged, and, therefore, the master confusion matrix contains the average confusions of participants across the whole range of acuities for each object.

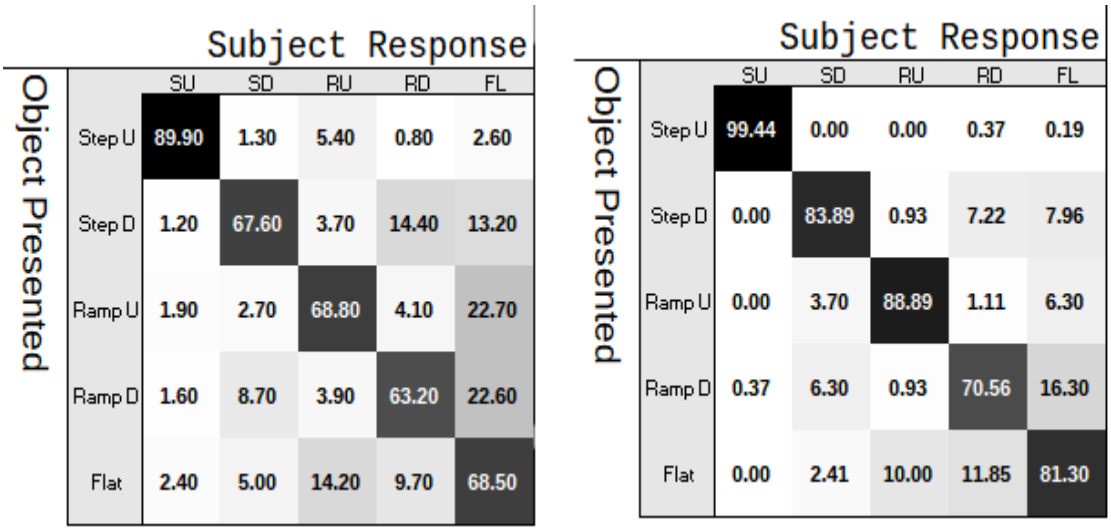
An analysis of similarity between the composition of two separately generated confusion matrices must demonstrate that the element by element ordinal properties of either confusion matrix is the same. If the confusion matrices represent sampling windows into consistent systems of information, it is likely that there are consistent patterns in the matrices themselves: e.g. flat is mistaken more often for anything else in both confusion matrices. If an element by element comparison of monotonic ordinal relations between matrices is stable, then it is likely that both matrices are sampling the same system of information. Therefore, it should be possible to compare the matrices by vectorizing them and computing a Spearman's rho between them. Likewise, a simulated null distribution can be created by randomizing one of the vectorized confusion matrices and thereby obliterating the ordinal monotonic relations within the system. Finally, this process can be repeated multiple times with a bootstrapping procedure to acquire a full distribution of both signal and noise.

Confusion Matrix Monotonicity. When confusion matrices exceed rank 2, it becomes difficult to glean meaning from them, but it is not impossible. The confusion matrix is oriented in such a way that the diagonal elements of the matrix represent correct responses to the trials: element [2,2] contains the number of times participants responded with object 2 when they were presented with object 2. Likewise, elements off the diagonal represent confusions; Off-diagonal elements of the matrix show how often participants were confused into making an error. The elements of the matrices then contain an implicit rank ordering to them; for example, element [1,2] and element [3,4] will always have the same monotonic relations between one another provided that the pattern of information the confusion matrix is sampling is stable.

It should be possible to compare between confusion matrices by comparing the rank ordering of monotonic relations of the elements themselves between matrices; if the patterns of monotonic relations between matrices is highly similar, a Spearman's rank order correlation should be close to 1. Furthermore, it should also be possible to simulate a null in that instance by scrambling the order of the elements prior to calculation of the Spearman's rho; such an act would randomize the monotonic relations of confusion elements and, thus, provide a sense of what noise would look like for the paradigm that produced the confusion matrices. At long last, it is possible to generate bootstrapped

Spearman's rho calculations by creating bootstrapped confusion matrices via randomly summing individual participant confusion matrices and calculating the Spearman's rho between the elements of the bootstrapped matrix and some other confusion matrix. After many repeated cycles of bootstrapping, the result is a distribution of Spearman's rho calculations that themselves represent the strength of monotonic relations within and between confusion matrix representations of observer behaviors. As alluded to earlier, a noise distribution can be created by randomizing the positions of elements in the bootstrapped confusion matrix prior to calculation of Spearman's Rho. Thus, the strength of similarity between observer systems can be further inspected by the distance between the means of the two distributions.

Results



Bochsler et al. Confusion Matrix Digital Confusion Matrix

Figure 5. On the right is the confusion matrix generated from the current experiment and on the left is the confusion matrix from Bochsler et al (2013) which represents the data gathered from participants with simulated low vision. The general pattern of participant responses between both studies is highly similar; however, the participants in the Bochsler study confused step-down stimuli with rams and flat surfaces more often than in the current study. The Bochsler et al participants also confused a ramp-up with a flat surface more often as well. The current experiment simulates a reduction in visual acuity without contrast loss. However, the previous experiment simulated Low Vision via pairs of vision obscuring goggles known as Bangerter foils. These foils attenuate visual acuity and contrast sensitivity simultaneously. This likely explains some of the general increase in accuracy in the current simulation.

Figure 5 shows the master confusion matrix generated by the digital object detection task. Objects presented are on the horizontal meridian and the vertical contains participant responses. As such, rows of the matrix should sum to 100, but column vectors

may not necessarily sum to 100%. The sum of a vertical vector's proportions reveals biases in participant responding. For example, participants responded with "flat" more often than a flat surface was presented. In comparison, "step up" had a near perfect response rate with little object confusions. Indeed, figure 2 shows a similar pattern in the confusion matrix from Bochsler et al (2010) where participants wore Low Vision simulating goggles.

Object classification patterns of subjects viewing 3D renders with simulated Low Vision were nearly identical to that of normally sighted participants doing the same task with the objects in a real space while wearing blurring goggles. Bootstrapped confusion matrices from the digital task were compared to the physical task's data with a Spearman's rank correlation of 0.9041, 99% CI [0.8672, 0.9410]. A simulated null showed a Spearman's rank correlation of 0.0079, 99% CI [-0.5164, 0.5006]. Figure 3 shows the histograms of the full matrix bootstrap distributions.

Object confusion patterns for subjects in the digital and physical tasks were also compared via the same bootstrapping procedure. The off-diagonal elements of the bootstrapped matrix from the digital task were compared to the off-diagonal elements from the matrix of the physical task. The Spearman's rank correlation was found to be 0.8745, 99% CI [0.8004 0.9486]. A simulated null of the off-diagonal element comparisons showed a Spearman's rank correlation of 0.0147 99% CI [-0.5690 0.5985]. Figure 4 shows the histograms from the off-diagonal element bootstrap distributions.

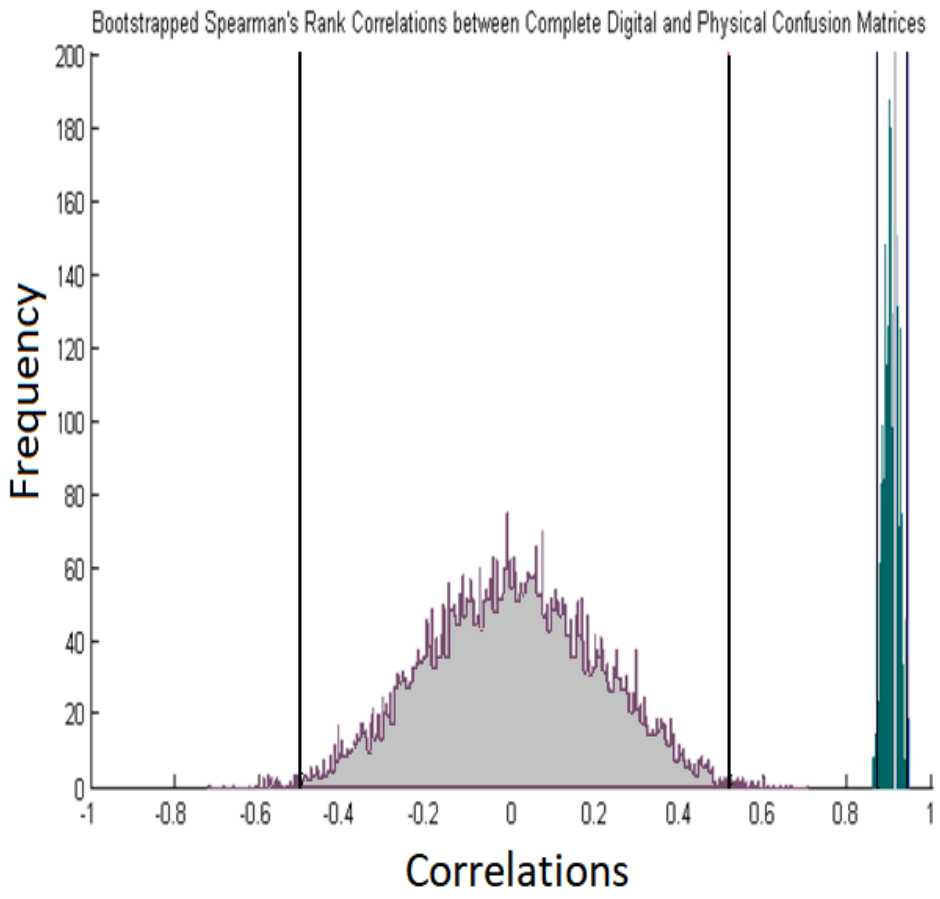


Figure 6. The display of bootstrapped confusion matrix calculations. The solid colored graph is the distribution of Spearman's rank correlations between the bootstrapped participant data and the confusion matrix from the physical experiment. The vertical black lines on the right side represent the 99% confidence intervals of that distribution. The gray vertical line is the actual Spearman's Rank correlation calculated between the data gathered in the digital task and the physical task. The solid and outlined graph is the noise distribution generated by first scrambling the order of elements of the bootstrapped participant data prior to calculating the Spearman's rank correlation. The vertical red lines represent the 99% confidence intervals of the noise distribution.

Discussion

In this study, normally sighted observers performed a digital version of an object detection task from Bochsler et al (2013) where observers had to identify pedestrian hazards under simulated Low Vision. In the digital version of the task, observers viewed images that had their spatial information attenuated in such a fashion as to produce simulations of six different Low Vision acuities, and these images were viewed under psychophysical conditions designed to match as closely as possible that of the real scene in terms of contrast and visual angle. The comparison of patterns of accuracies and confusions between the digital and physical tasks shows that both tasks produce

extremely similar response patterns. Therefore, the usage of 3D renderings as a substitute for construction of obstacle courses is methodologically adequate if one has taken care to reproduce the exact viewing angle in the laboratory as one would have had in the real scene.

Performance between the digital and physical tasks are not identical but they show highly similar patterns between elements of each tasks' confusion matrix: stair up responses were always greatest in both tasks, flat surface responses were always greater than the number of times displayed, downward steps were confused as downward ramps, and so on. Indeed, the average level of accuracy in the digital task is higher than in the physical task, but this is likely do to a difference in preparing participants for the task. In the physical task, participants were walked around the pneumatic device that formed the basis for all observed obstacles and were familiarized with the objects they were to identify. However, the digital task incorporated a guided training component due to its added mouse interface. It is possible that the training component familiarized the participants with the objects in such a way that inflated their accuracy; familiarity with a blurred object increases recognition accuracy (James et al., 2000). The similarities between the two tasks remain exceptionally high despite this, and, with the reduction in future training, it is likely that this similarity would only improve.

Most importantly, the confusions that participants made produced very similar patterns between the tasks, even though a great deal of image information is lost during tonemapping. The geometric nature of some obstacle remained as the defining feature for its detection; a ramp in the physical world remains a ramp in the digital world despite severe compression in luminance information. Truly, the dynamic range of an HDR image far outstrips the range of observable luminances that the eye can detect in a typical scene. The result of tonemapping is severe compression in the displayable range of information because there are luminances within the image that are simply too bright or dark to be displayed on a monitor, and the only way to display the image is to compress the entire range of luminances to fit within the displayable range of some display device. This compression effect produces a severe problem for reproducing real scenes with 3D renderings; if the detection of some pedestrian hazard relies on the contrast difference between differently oriented surfaces, it is possible that any loss of information could severely degrade detection performance. This study demonstrates that the careful calibration of display devices and mindful image processing of 3D renderings provides an avenue for using HDR images in the obstacle course laboratory.

Furthermore, the use of 3D rendering provides an opportunity for Low Vision obstacle detection experiments to be taken to new scenes which were cumbersome or impossible to build: subways, staircases, sidewalks, parking lots, fields, bus stops, and so on. With the usage of carefully tuned psychophysics, the real world can be transported into the laboratory. However, one must take care when reproducing some scenes in the lab. For example, nighttime scenes or extremely bright scenes would produce severe difficulty for adequate reproduction due to the nature of tonemapping. When one

tonemaps an image, information is lost, but this is typically done at the furthest end of the spectrum. A light bulb forms the bulk of the high end of the luminance range in an HDR image, but lightbulbs are seldom useful for detecting stairs outside of a movie theater. Thus, after tonemapping, the pixels that represent the lightbulb all become pure white; all information past some threshold would likely assume the same value. The same is true for extremely dark patches of images. However, this poses a problem for psychophysics because the compression of information may not match the luminance range of the display device. In other words, a monitor may not be able to display a pure black or white pixel and, thus, be unable to produce a one to one linear range of luminance values digitally as one would have physically even after tonemapping. This would be adequate if a comparison of the contrast ratios between surface orientations in a 3D render as compared to its physical counterpart produced a linear regression with a zero intercept; there was no bias in luminance contrast ratios between the scenes. However, this may not hold true for scenes that are at either extreme of a luminance spectrum and thus far beyond the display range of a typical research monitor. Despite this, most scenes do not occupy such extremes and observers would be unlikely to depend on that information because lightbulbs do not typically tell people the location of a sidewalk curb. However, it would still be prudent then to create scenes for which the luminance variation within them is not so large that they would be impossible to compress onto a monitor screen.

Much of the difficulty in transforming the real scene into the digital rendering lies in the display device: a monitor. The current experimental setup has a viewing distance of 13 cm, and this is close enough to the screen for observers to be able to see individual monitor pixels on a standard CRT. It is possible that such proximity to the display enhanced observant participants' accuracies; the differences between stimuli are slight. However, this is easily fixed with a larger display, and, indeed, future experiments using 3D renders whose purpose is to mimic real viewing conditions would do well to have the largest calibrated display available. It is also imperative that the display be linearly calibrated such that the contrast ratios between surfaces remains unbiased between scenes even when the scene luminances undergoes severe compression. The visual system allocates resources for depth processing in accordance to the statistics of the natural visual environment; simple luminance increments and decrements can influence perceptions of depth (Cooper and Norcia, 2014). Therefore, future works in this vein should endeavor to ensure that their displays of 3D renderings have as close to a 1-to-1 ratio with physical scene luminance as is possible.

The validation of 3D renderings creates many opportunities for Low Vision object detection psychophysics; scenes, objects, or scenarios that were once impractical to construct can be brought to analysis in the laboratory. This is particularly true for Low Vision research where the testing of obstacle avoidance is typically done with hand-made constructions. Whereas beforehand it was impractical, testing the visibility of a public work before it is constructed is now a realistic prospect. By using the methods outlined here, HDR 3D renderings of yet to be constructed spaces can be transported into the

laboratory. Indeed, obstacle courses are no longer the only avenue for testing Low Vision object detection patterns.

Chapter 3: Measuring Visibility of Geometry Edges in Low Vision Simulations

Background

The difference between seeing a sudden step down and missing it could be the difference between a boring Sunday afternoon and a broken hip. A "Watch Your Step" sign is a common sight in precarious pedestrian situations, but the sign only indicates the general area of concern and not the exact point where slippage may occur. Yellow curb delimiters are a more common and visually functional way of indicating steps or other changes in elevation (OSHA, 2017). Visual cues, such as curb delimiters, stair indicators, ramp outlines and so on, hint at civil engineers' intuitions that pedestrians are failing to see the liminal edges of navigational obstacles. Whether it be because of inattentiveness or inability, a pedestrian tripping onto a subway staircase is likely due to the innate challenge of delineating a scene into obstacles and smooth walkable floor.

The vertical and horizontal portions of a step in a subway staircase each form critical portions of the stair object: geometrical edges. In the case of single steps in a staircase, the edges of the steps are formed either by changes in surface normals of the rise and run of each step or by changes in depth because edges of the step geometry occlude background surfaces. If one were to change their perspective and look down the stairs, the stair faces would not be visible, but the geometry edges of the stair corners would still be in view, albeit slightly smaller because of a change in perspective. In image processing literature, edge detection is often performed on an image by looking for sudden changes in luminance; these locations in images often roughly correspond to the underlying geometrical edges of the scene that the image captured (Papari and Petkov, 2011). However, the lack of certainty in whether some apparent edge in an image really belongs to some true underlying geometry edge makes it fruitful to utilize two operational definitions when describing how an observer might infer the presence of geometry in a scene from an image: terms such as geometric boundaries or luminance boundaries.

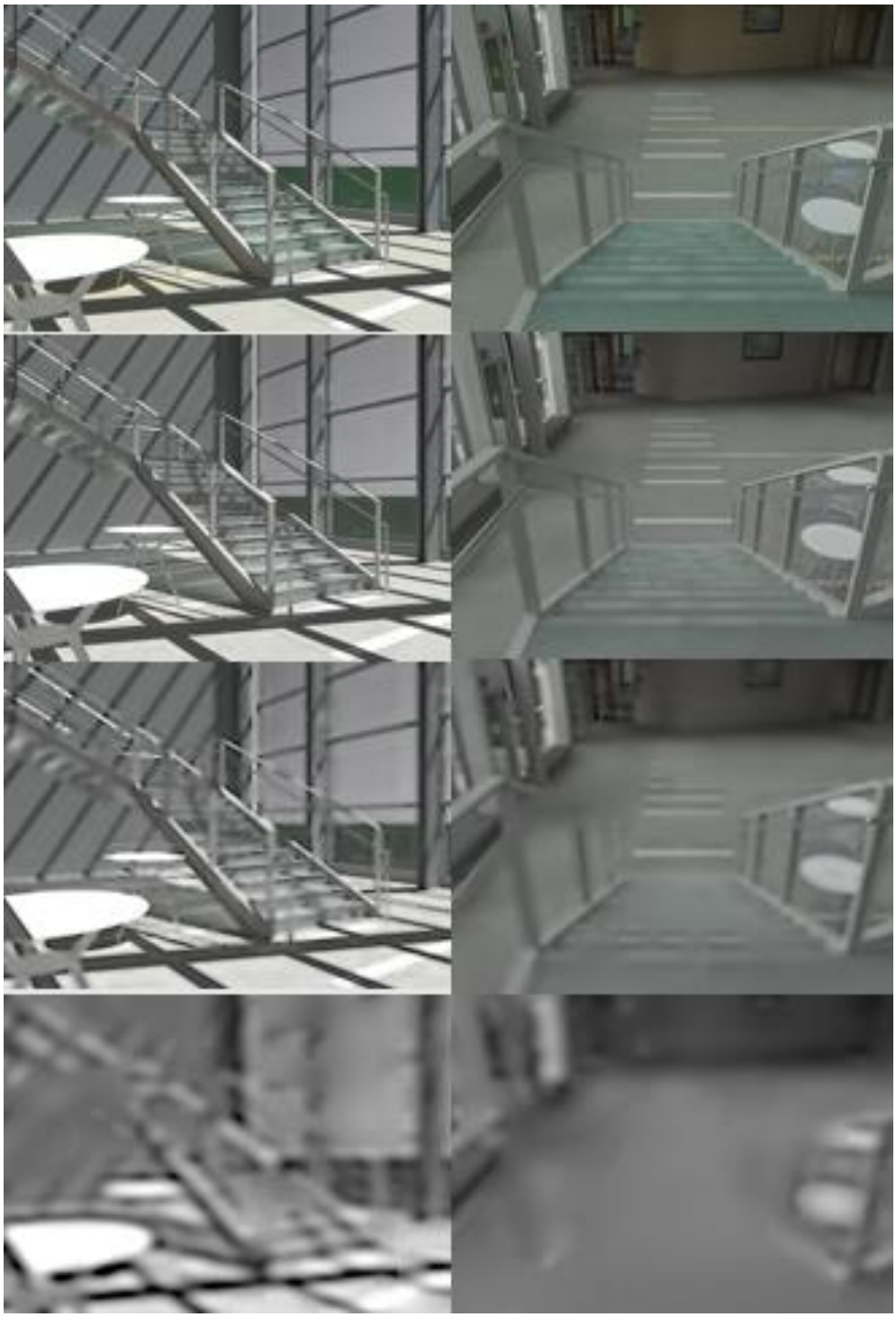


Figure 7. An illustrative example of situations where visibility of an obstacle's edges depends on both perspective and Low Vision simulation. Two images of the same scene are presented both unaltered and with three different Low Vision simulations: moderate, severe, and profound intensities. The images are of the same stair case but seen from different perspectives. As Low Vision simulations increase in severity, it becomes more

difficult to discern the contents of images but understanding a blurry scene can be harder when viewed with an unusual perspective.

From here onward, a geometric boundary is used to describe the physical boundaries that separate objects, delineate surfaces, and produce changes in depth. Luminance boundaries refers to image properties and contents that create boundaries in image pixels through changes in the luminance of contiguous lines of pixels that may or may not correspond to some underlying geometric edge. Such changes in the luminance of image pixels may be due to lighting and material changes as well as actual physical changes in the contiguity of surfaces. Sophisticated edge detection algorithms regularly find large amounts of edges in a scene yet only a small subset of these image edges correspond to the underlying geometry edges of the scene. (Canny, 1986; Bowyer, Kranenburg, and Dougherty, 1999). Observers utilize both texture and luminance-based changes in an image to infer the likely cause of an image edge in a scene (Vilankar et al., 2014). As acuity is reduced, texture information gets blurred out, leaving higher contrast luminance edges as the primary source of information for geometrical edges. In the interest of understanding the nature in which Low Vision observers detect or fail to detect the geometry of a staircase, it is imperative to know how a Low Vision observer interprets the blurry retinal image as a 3D geometric scene by parsing out how image edges in the retinal image are inferred as geometric edges.

A luminance boundary is a boundary formed in an image by abrupt changes in apparent luminance values in the image space. The presence of apparent changes in luminance across an image is not always a strong indicator of the underlying physical reality in an image; it is possible to trick the human visual system into seeing apparent changes in luminance that do not physically exist simply by organizing neighboring changes in surface reflectances to coincide with apriori expectations about object shading (Adelson, 2000). It is natural then to have the same intuition as civil engineers, when it comes to safety regulations, that a fall in a public space is a behavioral consequence of failing to see some luminance boundary caused by a geometric boundary. The designer of public space with an eye to safety highlights guardrails and stairs with the assumption that falls occur due to a failure of inference during passive analysis while walking through that space. The civic engineer or architect then has a second implicit assumption to their conscientious safety decisions that, if one were to take a snap shot of a public scene, the image edges of the scene that must be highlighted should be the ones that correspond to behaviorally relevant geometric edges.

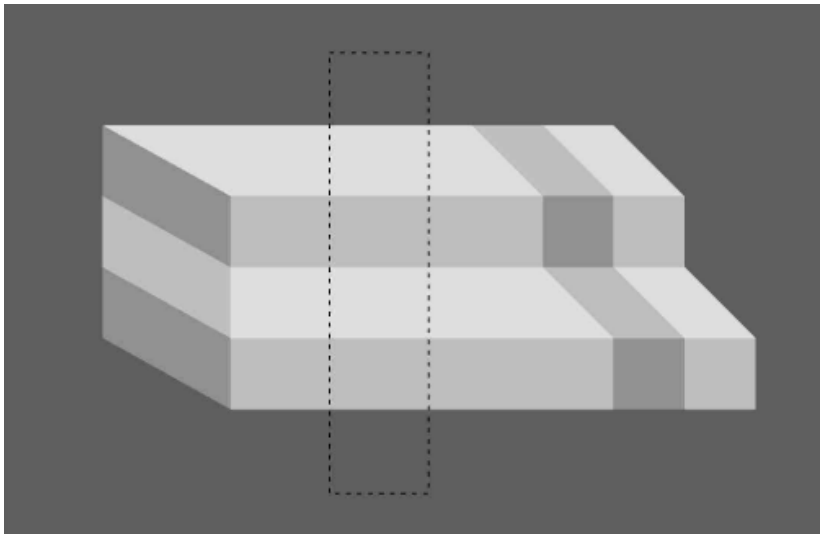


Figure 8. Adelson's Impossible Steps (Adelson, 2000), On the left side, the stripes in the image appear to be due to paint while, on the right, they appear to be due to shading.

Therefore, the civil engineer, the architect, or the interior designer often chooses to highlight the top lip of each stair in a staircase: the geometric boundaries separating changes in elevation that are assumed critical for pedestrian safety. However, Low Vision observers typically still have great difficulty segmenting scenes: parsing a scene into its constituent pieces. Low Vision observers' object recognition accuracy is enhanced more than normal observers' scores when the image edges of an object are pre-segmented (Bordier et al., 2011), but many public spaces remain visually inaccessible and accidents occur despite designer's efforts to highlight potential hazards. The mess on a table or the visual din of a subway can be puzzling to the Low Vision observer because it is not clear which image edges are relevant to the geometry of the scene and which are not: pre-segmenting the image space solves a sorting problem that is trivial to the normally sighted and taxing to the low sighted.

To sort useful from useless image information for navigation, it is likely that observers are inferring the presence of geometry boundaries by scanning the image space for strong luminance boundaries. To that end, the DEVA team has developed an algorithm that anticipates the visibility of geometry edges by using image edge information to aid in the design of accessible public spaces (Kersten, Shakespeare, Thompson, 2013; https://github.com/visual-accessibility/deva-filter). The algorithmic method anticipates the visibility of geometry edges by analyzing the power of luminance contrast at across multiple spatial frequencies in neighborhoods near geometric edges, and visibility is adjudicated by comparing the apparent level of contrast in image edge boundaries near geometry edges to a mathematically simulated Low Vision observer. The visibility of a contrast edge is rated by the distance of a geometry boundary pixel to its nearest luminance boundary pixel as estimated from a simulated Low Vision observer via their contrast sensitivity function: a key measure of contrast and spatial acuity both of which correlate with navigation behaviors (Legge, 2007). The result of the DEVA algorithm's analysis process is a throughput that gives visibility predictions on a point by point basis for all geometry edges in a scene

The predictive nature of the DEVA algorithm's image analysis process must be validated, if possible. It is necessary then to probe the predictive quality of measuring distance between image edges and geometry edges. The DEVA filter marks geometry boundary pixels as being less visible when the nearest luminance boundary pixel moves further away. It is predicted that a human derived visibility of geometry boundaries functions in the same manner. Human beings should be less likely to detect a geometry boundary pixel as luminance boundaries diffuse across the image space due to severity of Low Vision simulations.

However, the transformation of distance between geometry boundaries and luminance boundaries into visibility scores is arbitrary. Any visibility score derived from such a calculation is a scaling function that interprets the distance in a way that is meaningful to how human observers tend to fail to see geometry boundaries as a function of the nearest distance to some luminance boundary. To interpret that distance, two scaling functions are tested here: a reciprocal scaling factor and a Gaussian scaling factor.

In addition to testing the predictive veracity of distance scaling functions, it is also necessary to test the DEVA algorithm's predictions against globally derived image statistics such as high spatial frequency cut offs. An implicit assumption of the DEVA algorithm process is that its image processing throughput should outperform image statistic derivations in terms of predictive power due to the DEVA algorithm's simulation of Low Vision abilities which cannot be derived from an image of a scene alone. Therefore, it becomes critical that human data be tested on these simulations of Low Vision in the same way as the algorithm produces them; participants must trace objects in scenes under various Low Vision simulations. If the assumption that diffusion of luminance boundaries is the bellwether to geometry boundary inference, human tracings of obstacles in Low Vision simulations should track with the spreading of luminance boundaries across the image space as a function of Low Vision simulation severity.

To test the DEVA algorithm's visibility predictions, an edge labeling paradigm will be used; participants will be instructed to outline the edges of objects while viewing scenes under varying simulations of mild to severe acuity loss. Observers will be instructed to outline the edges of objects in three different categories: objects to avoid, regions of stairs, steps or ramps, and the floor wall boundaries of the space. The utilization of four different categories allows observers to make context sensitive decisions; under increasingly intense levels of Low Vision, participants might become overly parsimonious in tracings when not given explicit direction on what to label. By using pre-defined task relevant subsets of geometry boundaries, it is possible to enforce the DEVA algorithm to utilize subsets of scene geometry that are task relevant to each of the participants' object labelling tasks. It is then possible to directly compare a global image processing algorithm like DEVA to a more functionally oriented and task conscious observer like a human participant.

Methods

Simulation of Visibility. A general overview of the DEVA algorithm throughput can be seen in figure 9. The current build of the software tools for visibility computations are available and described in detail at https://github.com/visual-accessibility/deva-filter.

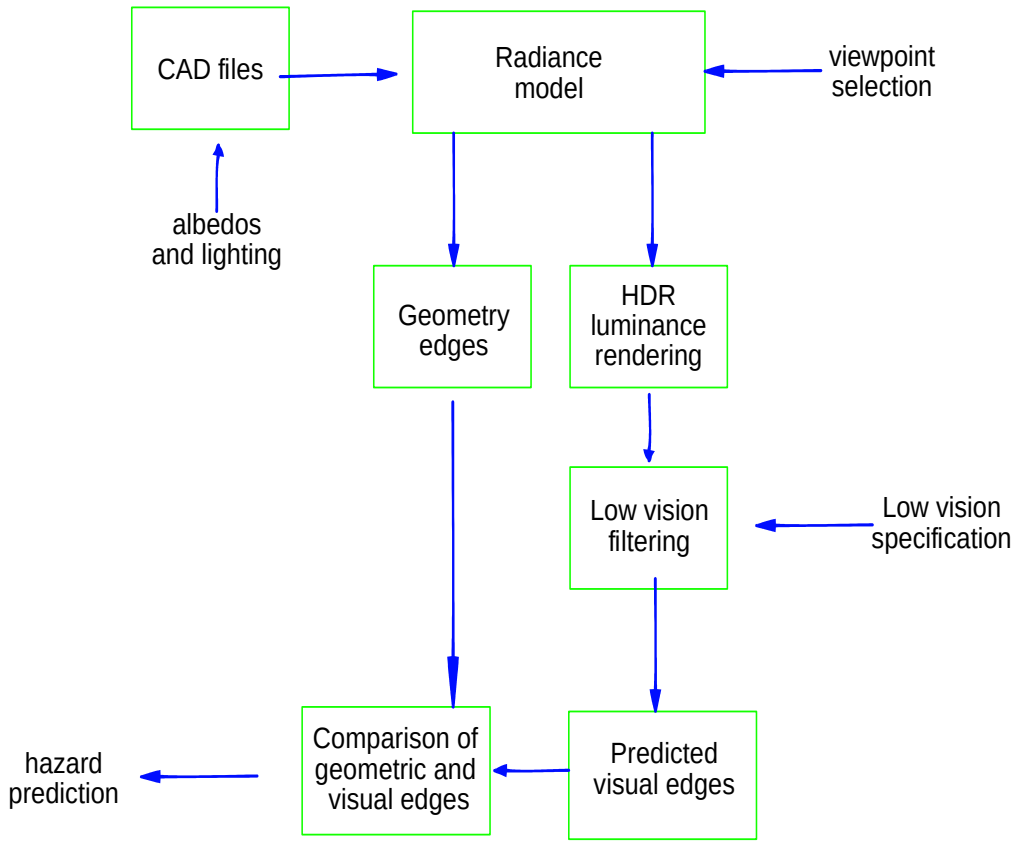


Figure 9: DEVA algorithm throughput process. The algorithmic begins with CAD files and Radiance models of a public scene. The algorithmic throughput then produces two separate streams of information: an HDR Low Vision simulation and a geometry edge mapping. The two come together to form an estimation of geometric edges predicted to be missed: hazardous edges.

The process begins by taking as input CAD file descriptions of the geometry, material, lighting and viewpoint of a scene and using Radiance (Ward & Shakespeare, 1998) to generate HDR luminance renderings of spaces that are under design. These renderings are photometrically accurate estimates of the pixel by pixel luminance values that would be measured in the real space from the specified viewpoint (figure 10, left panels). Radiance also outputs a list of depths and surface normals which are used to produce a complete geometry boundary map with respect to the viewpoint (figure 10, right panels).

Figure 10: Examples of geometry boundary maps built from HDR scenes and CAD files. The prototype algorithm employs a robust process that detects geometry boundaries in a scene. The geometry-boundaries algorithm selects nearly every geometry boundary in a scene: far more than any one participant is likely to select.

The HDR images of the scene are then filtered to simulate the elimination of visual structures with contrast below the detectability threshold specified by a clinical measure of acuity and contrast sensitivity thus simulating low visual ability. Filtering is done using a program (*deva-filter*) which accepts as arguments measures of acuity and contrast loss to create a CSF of the model low-vision observer (see Thompson et al., 2017). The *deva-filter* uses the CSF as part of an implementation of non-linear multichannel model based on an earlier model by Peli (Peli, 1990). The filter's output is qualitatively similar to the output of a simple convolution with a blur kernel, but in addition takes into account local variations in intensity across the image, and importantly leaves contrast above threshold unattenuated. The *deva-filter* has been validated with a clinical paradigm but not with behaviorally functional ones such as object recognition paradigms (Thompson et al 2017). The *deva-filtered* image is then run through a Canny

edge detector (Canny, 1986) to mark the locations of edges predicted to be visible (figure 11, right panel).

Figure 11: An illustrative example showing a scene, a severe Low Vision simulation, and the luminance boundary detection of that simulation image. The DEVA algorithm performs this step by using a Canny edge detector on the Low Vision simulation image.

The final step compares the geometry and visual edge maps to provide an estimate of visibility. The algorithm first measures the Euclidean distance in degrees of visual angle between each geometry boundary pixel and its *nearest* Canny edge boundary pixel. The assumption behind measuring this distance is that observers infer geometry boundaries in scenes from luminance boundaries, and that the distance between luminance boundaries and underlying geometry boundaries is an indicator of visibility. The rationale for measuring to the nearest edge is that the Canny edge detector (like most edge detectors) doesn't distinguish between edges caused by geometry vs. material or lighting (e.g. shadows); however, on average the Canny edge closest to a geometry edge is likely to be its effect.

Candidate measures of visibility, described below, are defined in terms of reciprocal or Gaussian functions of the Euclidean distance, each with a free scaling parameter. The experiments assess the extent to which these measures and scaling factors predict human tracing accuracy.

Stimuli. Scenes were high dynamic range (HDR) images generated and tonemapped in Radiance 3D Rendering (Ward & Shakespeare, 1998). The images were physically realistic renderings in terms of lighting and reflectance of surfaces. The images were displayed on a MultiSyinc E243WMi LCD monitor with a maximum brightness of 200 cd/m^2, a resolution of 1600 x 1200, and a refresh rate of 120 Hz. HDR images have a much larger range of luminance values than can be displayed on this current experiment's monitor, and to ensure accurate display of the HDR images with respect to keeping local contrast values linear across the image while also displaying the HDR image in a manner that is physically possible for the current monitor. The images were displayed in RGB with 256 color-mapping in a sRGB color space. The monitor was gamma calibrated such that the luminance values from the monitor were linear. When displayed, the images had an average luminance of 98.5 cd/m^2. The images were displayed with a resolution of 1200 by 1400 pixels.

The HDR images were physically realistic representations of a real public space: an atrium in the University of Indiana. For architectural design evaluation, each image is generated to represent the field of view and perspective of the observer. The images used here were originally designed to represent a range of realistic and typical views. As a result, the field of view for each simulated viewpoint varies from image to image; the visual angle across the simulated observer's field of view both horizontally and vertically changes from scene to scene. To ensure that participants in this experiment experienced the appropriate visual angle in each scene, it would have been necessary to have moved participants back and forth in the middle of the experiment because of the randomized presentation of stimuli. The stimuli had field of views that ranged from 70 to 50 visual degree spans horizontally and 50 to 45 visual degree spans vertically. Participants were seated at 30-cms from the screen. This viewing distance provided a 35 visual-degree vertical span for each stimulus. This distance was chosen to provide the largest visual angle presentation of stimuli without seating participants so close to the screen that they could count individual phosphors and while still being at least a factor of 2 within the largest vertical FOV span.

There were 44 HDR images in total, and the images were split into 4 different geometry boundary labelling groups: doors, floor-wall-ceilings, steps, and obstacles in the walking path. The doors and floor-wall-ceiling tasks had 10 images each, and the obstacles and steps tasks had 12 images each. The tasks overlap in content such that some images were shared between tasks. In total, there are 36 unique HDR renderings of public scenes.

The images were also separated by Low Vision simulation level: moderate (LogMAR 0.35, Pelli-Robson score 1.6, color saturation 75%), severe (LogMAR 0.75, Pelli-Robson score 1.0, color saturation 40%), and profound (LogMAR 1.55, Pelli-Robson score 0.5, color saturation 0%) simulations. The Pelli-Robson score is a measure of the log of contrast sensitivity. For example, a Pelli-Robson score of 2.0 indicates normal contrast sensitivity of 100 percent. A Pelli-Robson contrast sensitivity score of less than 1.5 is considered consistent with visual impairment.

Across acuity levels, there were 132 images: 44 at each level of Low Vision simulation. The HDR images were processed by the prototype DEVA algorithm to produce mathematically accurate simulations of low visual ability; this primarily involved a reduction both in spatial frequency content and contrast power in images that increased in reduction with more and more profound simulations of Low Vision (Thompson et al., 2017).

Participants. 30 undergraduate students from the University of Minnesota took part in the experiment. The experiment was approved by the Institutional Review Board of the University of Minnesota. Each participant completed a single experimental session that lasted anywhere from one to two hours. Each participant completed all experimental tasks at a single acuity level only; there were 10 participants in each acuity level. After the experiment was completed, participants were rewarded with extra credit points for the time they spent participating. Each participant had corrected to normal or normal vision.

Experimental Procedure. Prior to the experiment, participants were briefed on the demands of the task, the nature of Low Vision, and that they would be using a mouse to

identify and trace the contours of objects in images that simulated Low Vision. Participants then performed a short training session that measured their motor error while tracing a straight line. The training session oriented them toward how to use the experimental software to trace object contours, erase selections, and move on to subsequent stimuli.

During the training session, participants were taught the instructions that they would be using to label geometry boundaries in task blocks that would follow. The experiment contained four separate geometry boundary labelling tasks sorted into separate task blocks: doors, stairs, floor-wall-ceilings, and obstacles in the walking path. In each task block the participant was instructed to only label the locations they believed corresponded to some underlying geometry boundary location and to only label geometry boundaries that were task relevant. For example, if they were currently labelling geometry boundaries in the doors task, the participant should only label those parts of the image that they felt corresponded to underlying door geometry. The instructions are as follows for each of the tasks:

- *Doors: "Outline the edges of doors and door frames."*
- *Floor-wall-ceilings: "Draw the edges of the space itself. Outline the points where the floor to wall, wall to wall, and ceiling to wall sections connect."*
- *Steps: "Outline the edges of the first 3 stairs you would encounter and also highlight any handrails within grabbing distance of those stairs. In addition to this, if a handrail is present, draw a line along the length of that hand rail to the point where you can no longer see the handrail"*
- Obstacles in the walking path: *"Outline the outer boundary edges of objects you could potentially run into, and also do your best to highlight the edges to the interior of this boundary. For example, if you see a box, outline both the contour of this box and also the locations that form its "interior" corners."*
- *"… Only label those portions of the image that are actually visible to you."*

After the training session, participants were randomly sorted into one of three acuity conditions, either moderate, severe, or profound vision loss, and randomly began one of the four mentioned geometry boundary labelling task blocks. Participants completed all four tasks in random order, and each task block contained only task relevant images which were also randomized for each participant. Participants traced the locations of geometry boundaries only once for each image within each block. Prior to beginning a new task block, participants were reminded of the instructions for the relevant task. During the experiment, participants sat 30 cm from the screen.

By the end of the experiment, participants had traced the locations they believed to correspond to task relevant geometry boundaries in 44 HDR images at a single simulation of Low Vision. After the completion of their tracings, the participant was debriefed and interviewed about their experience.

Expert Geometry Templates. Prior to analysis, it is necessary to create subsets of geometry boundaries that function as baselines for accurate geometry boundary detection; these subsets of geometry boundaries are known as expert templates. Each expert geometry template is unique to the scene and task from which they are created and each one contains all relevant geometry boundaries for some scene at some task. As seen in figure 10, many images of public scenes contain a great deal of geometry boundaries, but only a small subset of those boundaries are relevant for an ambulating pedestrian at any given point in time. Without task relevant geometry boundary labeling instructions, observers would have to label every geometry boundary they believe exists in an image: highly impractical and time consuming. Furthermore, subsets of geometry boundaries can be used by the prototype algorithm to create visibility scores for only those geometry boundary pixels. It is therefore possible to create a standard of task relevant geometry boundary accuracy that functions both for human participants and the automated prototype.

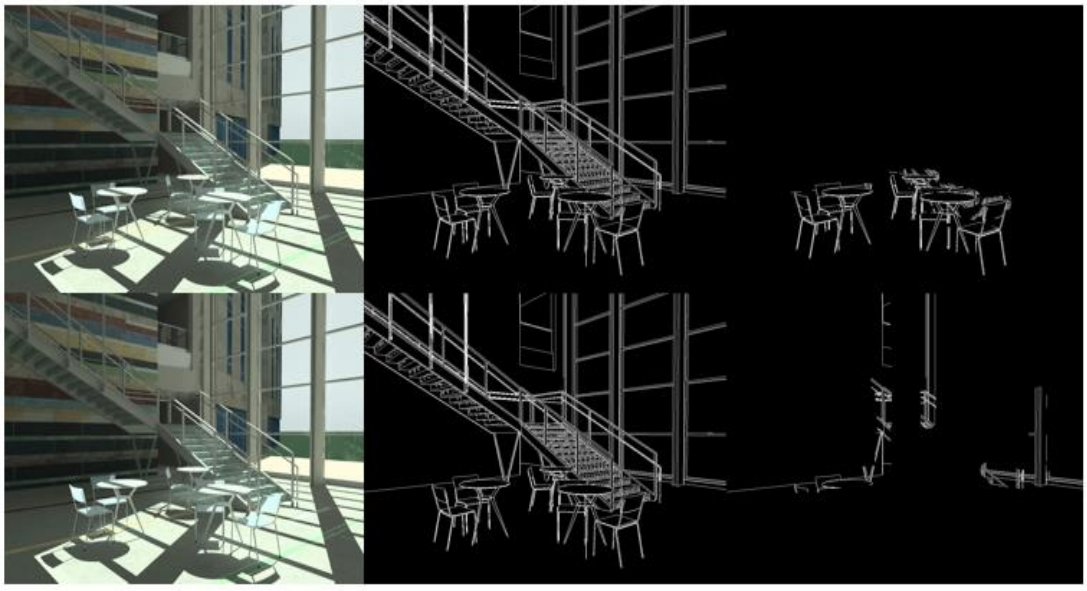


Figure 12: An example of a scene, a complete set of geometry boundaries for that scene and an expert template. In the first row, one can see the expert template for obstacles in the walking path, and the bottom row shows the same images for the scene and geometry but with the expert template for the floor-wall-ceilings (FWC) task. The expert template has few extraneous details besides chairs and tables while the FWC template has extraneous details that surround the geometry pixels of walls and floors. Tracing of geometry boundaries by participants will always capture these extraneous pixels because tracing accuracy is calculated with lateral spans across the image space in a manner like expert template creation.

To create expert geometry templates, three experienced psychophysics participants (a professor, a post doc, and the author of this thesis) all independently traced non- blurred versions of scenes for each task such that they captured what they believed were all relevant portions of the scene that corresponded to doors, steps and so forth. An expert selection of geometry was created for each expert observer by sorting through

scene geometry edge pixels and running a detection for any expert labelling within a half visual degree of that pixel. If a geometry edge pixel had an expert selection within a half visual degree, it was included in that expert's template. The creation of the final expert template for any one scene is the combination of all three experts' opinions which can be done with a simple set of rules. One can either create the union of expert opinions, e.g. all expert selections are included, or one can create an intersection rule where a plurality of experts must agree upon any single geometry edge pixel to include in the final template. Because of the relative concision in expert response and tendency for experts to agree very often, the union templates are used for all analyses here.

Analysis Procedure. Visibility scores are produced in a multi-step process that begins by taking the relevant expert template for a scene, and then, for each geometry boundary pixel in the expert template for that scene, measuring the Euclidean distance between a geometry boundary's pixel location and the nearest input pixel, either a luminance boundary pixel from the prototype process or a tracing pixel from the participant, and, finally, scaling that distance with some visibility metric The utilization of subsets of geometry boundaries as expert templates allows one to enforce task specific analyses of visibility. Once every geometry boundary pixel has an associated Euclidean distance measure, it is necessary to interpret those distance scores.

In the final stage, visibility scores are created by interpreting the distance score between each geometry boundary pixel and their nearest neighbor input. At this early stage of validation, any chosen metric need only produce lower visibility scores when the geometry and image edge pixels are far apart and high scores when they are closer together. Two simple distance metrics are employed here to create visibility measures as a function of distance: a reciprocal scaling metric:

$$\text{Visibility} = \text{scale} / (\text{degrees per pixel} * \text{input} + \text{scale})$$

and a Gaussian scaling metric

$$\text{Visibility} = e^{\wedge} (-1 * (\text{degrees per pixel} * \text{input} / \text{scale}) ^{\wedge}2)$$

where input is the Euclidean distance between geometry boundary pixels and input pixels, scale is a free parameter, and degrees per pixel is a conversion factor that turns the Euclidean distance between pixel locations into visual degrees. This conversion factor varies very slightly from scene to scene and is determined by the field of view of the simulated observer when the CAD model specifies a viewpoint to be rendered into an HDR image. The reciprocal scaling metric is one that creates visibility scores that steadily drop to 0 as the nearest image edge input moves away from the relevant geometry pixel, and the Gaussian scaling metric produces visibility scores that remain strong at interim distances before quickly advancing to 0 with greater and greater distances. Both metrics have a free parameter that can either be set to any vale which allows for an adjustment of fit between algorithm visibility scores and human participant visibility scores.

Image statistics. The argument for the DEVA filter contains the implicit assumption that one can more strongly predict the visibility of geometry by measuring the relative distance between geometry boundary pixels and nearest neighbor luminance boundary

pixels than by deriving global image statistics for a scene. If this were not the case, the DEVA filter's predictions would be obviated by a simpler process that calculates global scene statistics and predicts visibility by knowing the delta of visibility degradation as a function the decreasing parameter space of those image statistics. It is then imperative to test the DEVA filter's predictive capabilities against globally derived image statistics such as the highest spatial frequency element of a scene as a function of lowest visible contrast or the root mean square contrast of a scene's luminance values. These two statistics are chosen before others due to their being an essential element of the DEVA filter's simulation of Low Vision CSF.

The first globally derived image statistic is the high spatial frequency cut-off of a scene as a function of some criterion contrast in the amplitude spectrum of the Fourier transform of that scene. This statistic is chosen because it directly reflects the assumption that Low Vision can be simulated by mathematically deriving a CSF, parameterizing lateral and vertical reductions in that CSF, and using those parameters to attenuate spatial frequency and contrast in the image space. Any simulated Low Vision observer from such a process would have a maximum spatial frequency at some contrast level that they would be sensitive to. Thus, it is possible that a similar statistic derived from scenes might more accurately reflect visibility of geometry edges if the primary driver of behavior is merely the global image statistics of the scene rather than how the observer interprets local image properties such as image edges.

High spatial frequency cut-offs of the power spectrum of scenes are calculated as part of a multi-step Fourier analysis process. First, the amplitude spectrum of a scene is found. This is done by taking the Fourier transform of an image which produces the amplitude spectrum of the scene with complex numbers. The absolute value of the squared values of that analysis produces the power spectrum of the scene. The second step is to take the power spectrum and zero-point the graph by shifting zero frequency elements to the center of the power spectrum. The goal of this analysis is to summarize the cut-off in the amplitudes across all spatial frequency orientations in the image space, and, to do that, it is necessary to average across orientations in the power spectrum. Therefore, the third step is to interpolate an averaging function across the spectrum in polar coordinates. This produces the average contrast power of spatial frequencies across the image space. The fourth step is to calculate a cut-off frequency for some criterion level; the choice of criterion is arbitrary and here the assumption is that the visual observer cannot be sensitive to some contrast power below 0.001. The highest spatial frequency cut-off in a scene is then solved for by solving a linear regression for the interpolation function and returning the expected frequency for the arbitrary contrast criterion.

RMS analysis. Contrast reduction is the second part of the prototype algorithm's Low Vision simulation process. Contrast is the difference in luminance between a pixel and a reference luminance. The contrast of a sinewave grating is typically determined with a simple contrast function such as Michelson's contrast: the difference in luminance between brightest and darkest spots divided by the sum of luminances from those same spots. However, simple measures of contrast are not meaningful for entire complex images. To measure the relative contrast in HDR scenes, one needs a measure that spans the entire image space and gives contrast as a function of the average deviation in

luminance across that space. For that purpose, one can employ a definition of contrast known as the root-mean-square of contrast luminance levels. RMS contrast is the standard deviation of pixel luminance in an image divided by the average luminance of the entire image. For analysis purposes, RMS contrast for each scene in the experiment is compared to the average participant generated visibility for those scenes.

Jaccard Index Analysis: A Jaccard index is the intersection divided by the union of two sets; the number of shared elements over the total number of elements. It is an index that shows the similarity between sets of information, and it is employed here on a geometry boundary detection comparison between human tracings and luminance boundaries. A second geometry boundary detection analysis was performed to find any inputs, either tracing pixels or luminance boundary pixels, within stepwise visual angle distances of geometry boundary pixels. The steps were in quarters of a single visual angle until 1 visual degree and then half visual angle steps until 3 degrees across the image from each geometry pixel. Geometry pixels were rated on a binary scale of visibility for each visual angle distance where they were rated as visible if an input was within the specified visual angle radius. This was repeated for each visual angle distance for each geometry boundary pixel in an expert template for each scene in every task at all three acuity levels. Visibility maps of geometry boundary pixels in the expert templates were separately generated for both luminance boundary inputs and participant tracing inputs. Jaccard indices were generated at each visual angle by detecting whether both the luminance boundaries and participant tracings were within the visual angle radius for each geometry boundary pixel in the expert template for that scene. If both luminance boundaries and participant tracings were within a specified radius for a geometry boundary pixel in the expert template, the Jaccard index increased.

Results

Figure 13 Demonstration of scene, simulation, and participant data. The left image is a door scene from the doors task, the middle image is a simulation of profound Low Vision, and the right image is an example of a participant's response to finding doors in that profound Low Vision simulation image.

30 participants traced 44 simulations of Low Vision scenes which produced a total of 1320 data images. These data images are broken down and analyzed both within and across acuity and within or across all tasks. Average participant performances were created by calculating visibility scores for each scene. Pearson correlations were calculated by pairing the average participant visibility score for a scene with the predicted

DEVA visibility score for that scene for each of the tasks. The correlations were calculated across acuity levels and the correlation process was repeated for both scaling metrics. The distance metrics used to produce visibility scores utilize a free parameter. The initial correlations reported here are with a scale factor of 1; this choice is arbitrary and correlations spanning the free parameter space are reported as well.

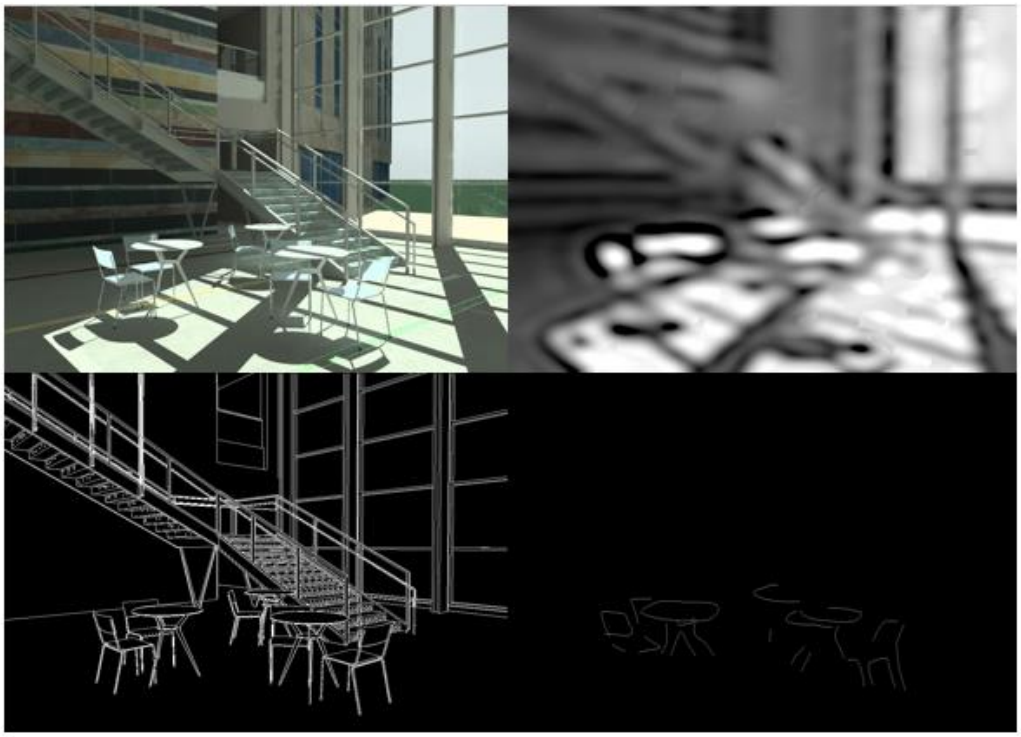


Figure 14: An example of an image of a scene, a simulation of profound Low Vision, a complete mapping of geometry boundaries for that scene, and a participant's selection of obstacles in that scene during a profound vision loss simulation.

For the reciprocal scaling metric, correlations within tasks and across acuities between average participant visibility and DEVA visibility were significant for all tasks ($r(28) = 0.52$, $p < 0.01$ for door tracings, $r(28) = 0.38$, $p < 0.01$ for floor-wall-ceiling tracings, $r(28) = 0.67$, $p < 0.01$ for obstacle tracings, and $r(28) = 0.57$, $p < 0.01$ step tracings). Correlations were calculated in the same manner with visibility scores generated by the Gaussian scaling metric and correlations for all tasks by the floor-wall-ceilings task were significant with the Gaussian metric ($r(28) = 0.50$, $p < 0.01$ for door tracings, $r(28) = 0.30$, $p = 0.11$ for floor-wall-ceilings, $r(28) = 0.61$, $p < 0.01$ for obstacle tracings, $r(28) = 0.52$, $p < 0.01$ for steps tracings).

Task	Reciprocal	Gaussian
Doors	r(28) = 0.52, p < 0.01	r(28) = 0.50, p < 0.01
Floor wall ceilings	r(28) = 0.38, p < 0.01	r(28) = 0.30, p = 0.11
Obstacles	r(28) = 0.67, p < 0.01	r(28) = 0.61, p < 0.01
Steps	r(28) = 0.57, p <0.01	r(28) = 0.52, p < 0.01

Table 1: Correlations of average participant visibility scores per scene to algorithm prediction visibility scores across acuities in single tasks.

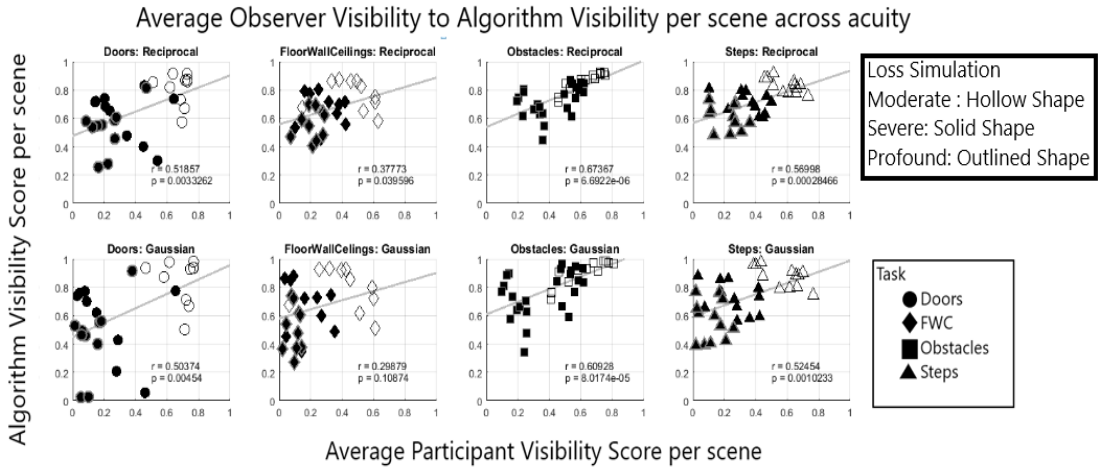


Figure 15: Correlations of average participant visibility scores per scene to algorithm prediction visibility scores across acuities in single tasks. Across acuity levels, the prototype algorithm predicted participant geometry detection behavior well, but the prototype algorithm tends to overestimate visibility of scenes on average.

Correlations were also produced across tasks and across acuities for both scaling metrics as well. This was done in the same manner: paired average participant scores in a scene and task to the DEVA prediction with that scale metric. The DEVA filter using the reciprocal scaling metric showed a strong correlation with average visibility of scenes across acuity levels ($r(130) = 0.56$, $= p < .0001$). Likewise, the DEVA filter using the Gaussian scaling metric had a correlation strength that nearly matched the reciprocal metric visibility correlation ($r(130) = 0.51$, $p < .0001$).

Correlations were produced within each of the three acuity levels and across tasks for both distance metrics. All Pearson's correlations within all acuity levels and across task between DEVA filter predicted visibilities and average participant visibilities were insignificant using the reciprocal distance function and Pearson's correlations with the same comparison but with values derived from a Gaussian distance metric were also insignificant. Linear analyses across tasks and within acuities between participant and prototype geometry boundary visibility shows no linear relationship. However, the relationship is preserved and strong across acuity.

	Metric	
Acuity	Reciprocal	Gaussian
Moderate	r(42) = 0.15, p =0.34	r(42) = 0.09, p = 0.58
Severe	r(42) = 0.13, p =0.39	r(42) = 0.15, p = 0.343
Profound	r(42) = 0.20, p =0.193	r(42) = 0.25, p = 0.095

Table 2: Correlations between average participant visibility scores per scene and algorithm predicted visibility scores within single acuity levels collapsed across tasks.

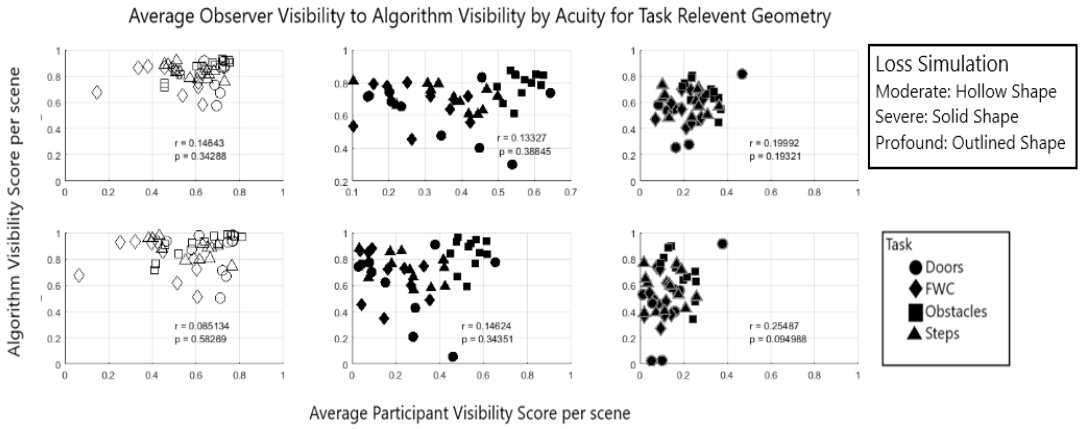


Figure 16: Correlations between average participant visibility scores per scene and algorithm predicted visibility scores within single acuity levels collapsed across tasks. The prototype algorithm's visibility predictions do not match participant geometry detection within single levels of Low Vision simulation.

Linear analyses across tasks and across acuities found that a significant proportion of variance in visibility scores was explained by the prediction algorithm's expectation that a reciprocally scaled distance to luminance boundaries should predict visibility of geometry boundaries in scenes ($F(1,130)=58.4$, $p < 0.001$) with an R^2 of 0.31. A Gaussian scale factor of visibility scores was also able to explain a significant proportion of participant visibility scores ($F(1,130)=46.4$, $p < 0.001$) with an R^2 of 0.263.

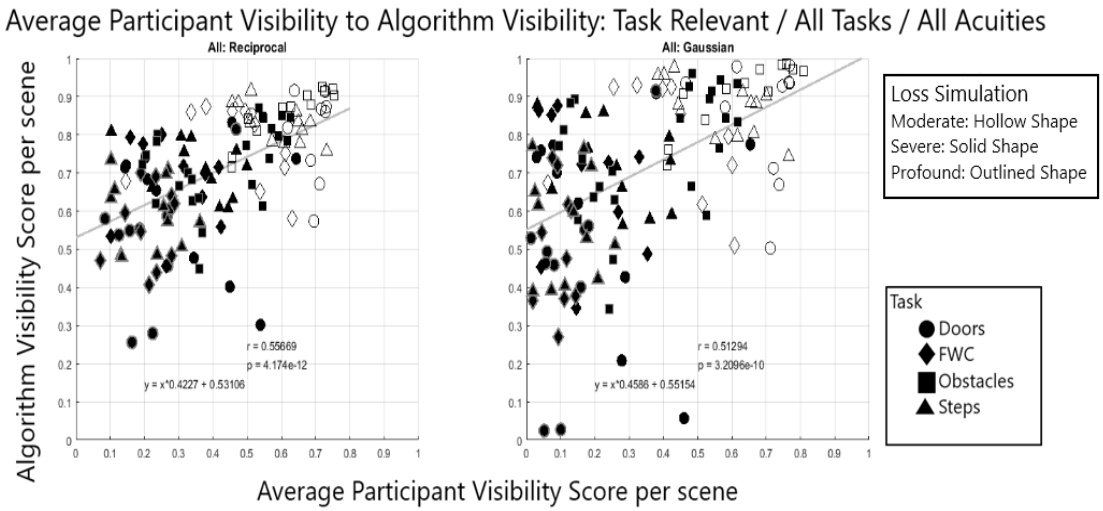


Figure 17: Correlations between average participant visibility scores per scene and algorithm predicted visibility scores across acuities and across tasks. Both reciprocal and Gaussian metrics performed at about the same rate across tasks and acuities.

A linear analysis was performed on the residual values of the linear regressions performed for both distance metrics on visibility scores across acuities and across tasks. The analysis found that both measures were highly correlated with one another ($F(1,130)=529$, $p < 0.001$, $R^2 = 0.976$), but with an average slope weighted towards the Gaussian scale factor. The comparison of residuals found a slope of 1.4855 ($t = 72.754$, $p < 0.001$) indicating that the Gaussian distance metric tended to both over and under estimate visibility scores on a scene by scene basis as compared to the reciprocal distance metric which has less prediction error on average.

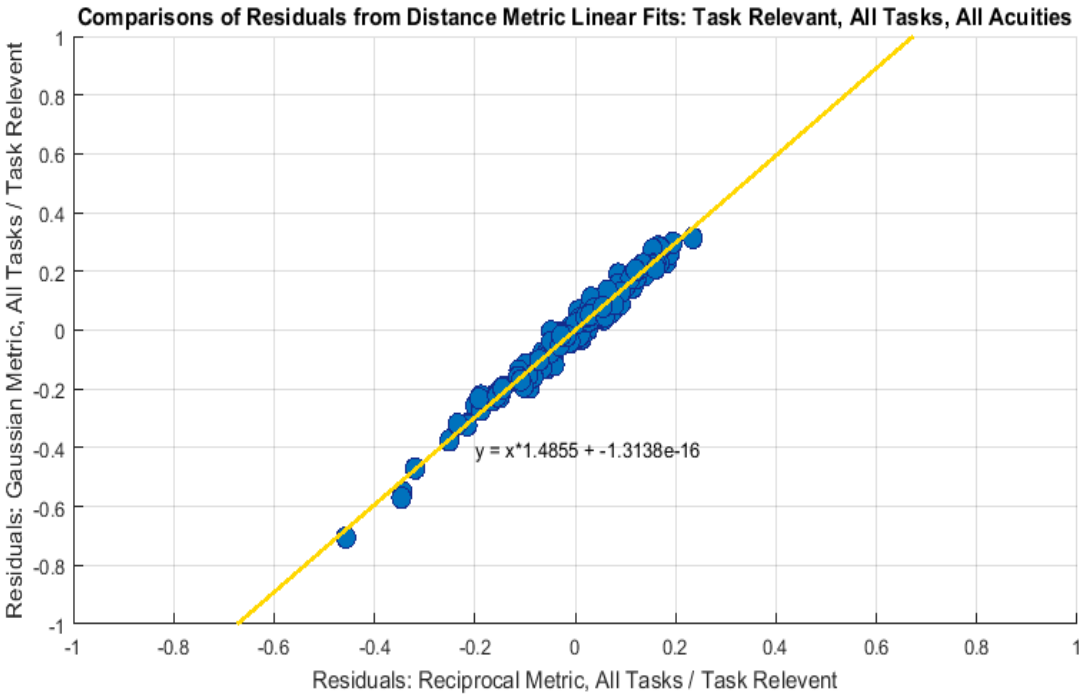


Figure 18: Linear analysis of residuals from the linear regressions of gaussian and reciprocal fits across all tasks and acuities. The linear analysis of residuals has a slope weighted towards the Gaussian metric. This indicates that the Gaussian metric tends to produce slightly larger prediction errors.

Visibility scores were generated across multiple scale factors for both distance metrics. Scale factors were 0.25, 0.5, 1, 2, 4, 8 inserted as the free parameter in the reciprocal or Gaussian distance metric equations. Correlations were then generated within task and across acuity for all scale factors with both distance metrics. The results can be seen in figure 17. On average the correlation between prototype algorithm visibility prediction and average participant visibility on scenes in all four tasks tended to increase as scale factors decreased with the notable exception of the obstacles task which reaches a peak similarity at a scale factor of 1.

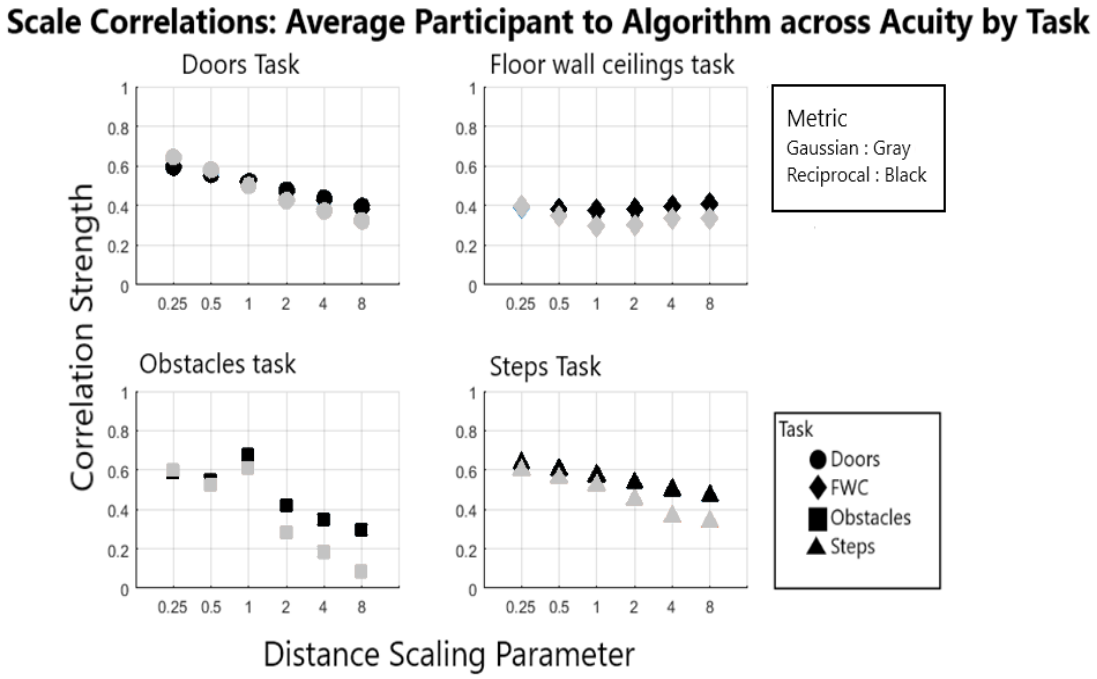


Figure 19: Analysis of scaling factors for both distance metrics. As the scaling factor approaches 0, correlations between participant and prototype visibility scores tends to improve. The exception is the obstacles task where correlations peak with a scale factor of 1.

Image processing and Jaccard analyses

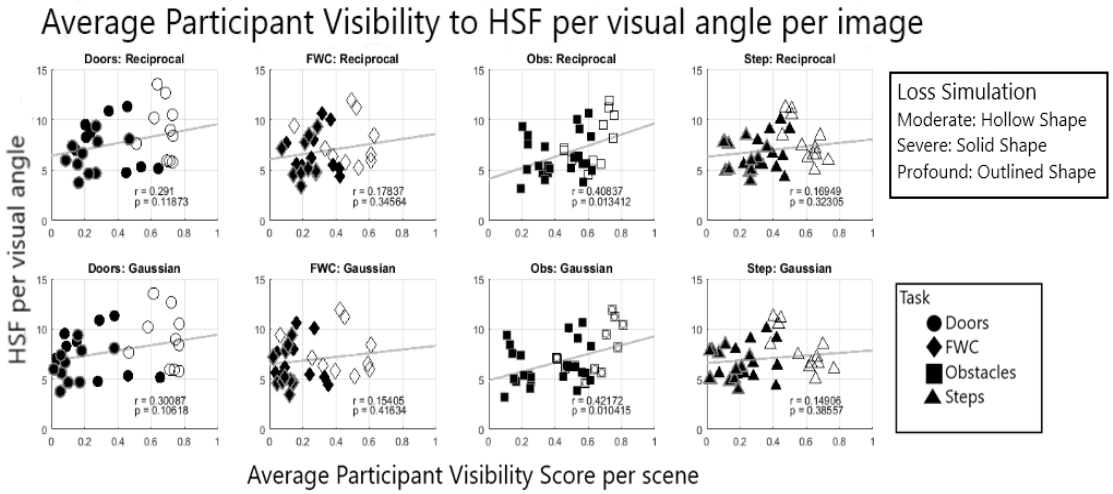


Figure 20. Average participant visibility scores per scene compared to the high spatial frequency value derived from that scene. HSF is not predictive for any task other than obstacle detection. This is likely due to obstacle detection being a local task requiring fine spatial detail whereas the locating of a staircase can be done with only global scene context.

Linear analysis across acuities within each task for both distance scaling factors indicated that high spatial frequency for a scene was only predictive of obstacle visibility and not doors, floor-wall-ceilings, or steps. The high spatial frequency of a scene was predictive when one used either reciprocal ($F(1,34)=6.8$, $p=0.0134$, $R^2 = 0.167$) or Gaussian distance metrics ($F(1,30)=7.35$, $p = 0.0104$, $R^2 = 0.178$). An exploratory analysis of the average and spread of HSF of images across all tasks in all acuity levels can be seen in figure 21: in all tasks, variance remains high while average HSF decreases.

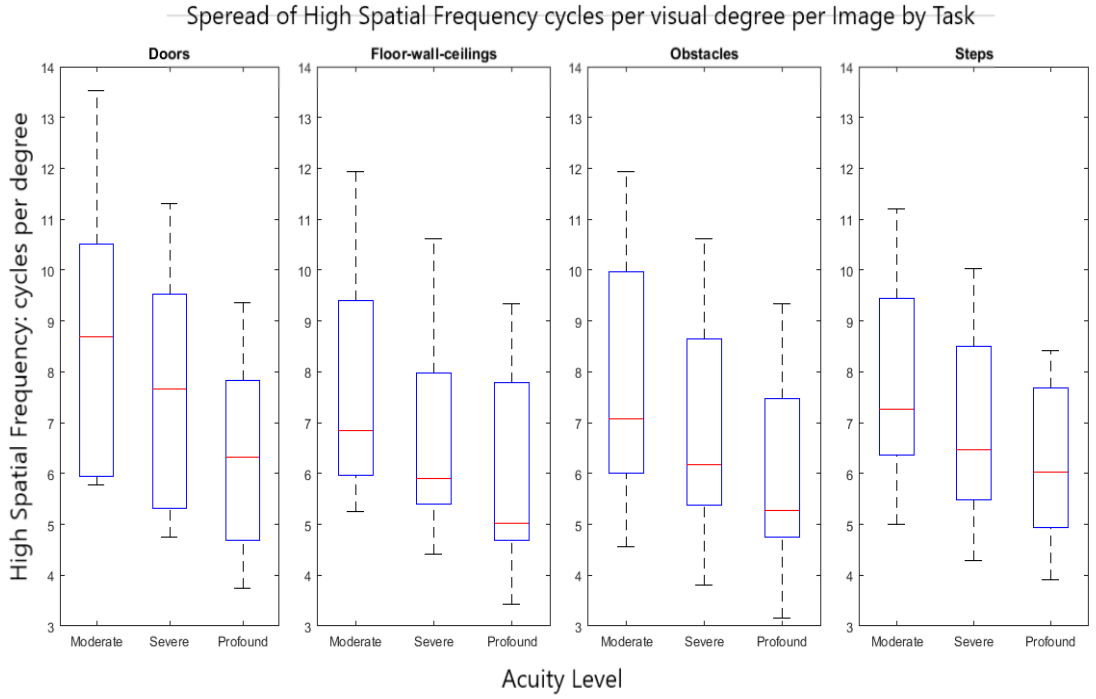


Figure 21: A boxplot analysis of the average and spread of high spatial frequency in cycles per degree for images across all four tasks and in all acuity levels. As the average HSF decreases with acuity Low Vision simulation intensity, the spread of HSF across the set within an acuity level remains large in all tasks.

An Analysis of root-mean-square contrast to average participant visibility per scene returned no significant results: RMS was not predictive of participants' ability to detect geometry boundaries in any of the tasks. Correlations between the average participant visibility score for each scene in each task across acuity simulations and the RMS contrast of those scenes returned insignificant values for all comparisons regardless of whether visibility scores were generated with the reciprocal or Gaussian scaling factor.

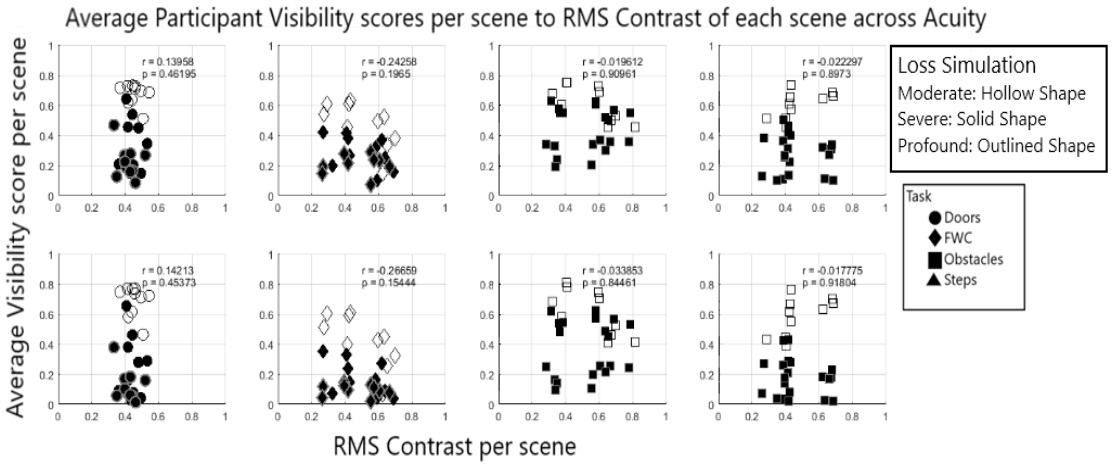


Figure 22. Graph showing average participant visibility scores to the RMS contrast per scene across acuity and in each task separately. Correlations in each individual task across acuity between the average visibility score and the RMS contrast per scene were insignificant in all tasks regardless of distance metric used to derive visibility scores.

Stepwise Jaccard Index Analysis Results. Jaccard indices calculated for geometry boundary detection agreement between the prototype algorithm and participants at stepwise radii found increasing agreement between the two at all acuities with growing detection radii. The 95% confidence intervals for average Jaccard Indices across scenes for each task at every stepwise detection radius were calculated and can be seen in tables 3 to 6. Almost every Jaccard index confidence interval in moderate and severe acuity simulations overlaps at almost every single detection radius in most tasks, but the confidence intervals for Jaccard indices gathered at profound vision loss simulations do not overlap with any of the intervals from either of the other two simulations. There is strong disagreement across tasks over geometry boundary detection on a pixel by pixel basis between participant tracings and prototype algorithm luminance boundaries at profound vision loss simulations.

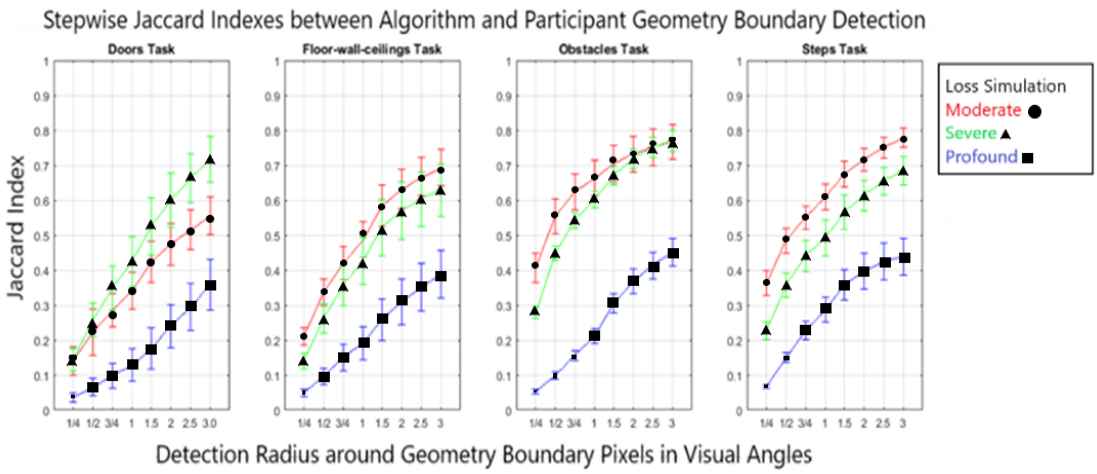


Figure 23**.** Stepwise Jaccard Indices showing agreement between luminance boundary and participant tracing detection of geometry boundaries at expanding radii per geometry pixel. A Jaccard index increases when each geometry boundary pixel is separately found to have a mutual input pixel within a specified detection radius. Detection radii expand in quarter steps from 0.25 visual angles to 1 visual angle and then expand in half visual angle steps from 1 to 3 visual angles. The prototype algorithm and participants tend to strongly disagree during simulations of profound Low Vision, even at very large detection radii. This is due to participants' tendencies to not respond at very profound vision loss simulations.

	Doors-Moderate	Doors-Severe	Doors-Profound
0.25	(0.11-0.23)	(0.12-0.21)	(0.02-0.06)
0.50	(0.17-0.36)	(0.22-0.36)	(0.04-0.12)
0.75	(0.22-0.46)	(0.3-0.48)	(0.07-0.17)
1.00	(0.27-0.53)	(0.37-0.58)	(0.09-0.23)
1.50	(0.35-0.63)	(0.46-0.71)	(0.13-0.31)
2.00	(0.4-0.68)	(0.54-0.77)	(0.19-0.38)
2.50	(0.44-0.72)	(0.61-0.81)	(0.24-0.45)
3.00	(0.48-0.76)	(0.67-0.86)	(0.3-0.52)

Table 3. The doors confidence intervals of Jaccard indices at detection radii from ¼ degree visual angle to 3 degrees visual angle around geometry boundaries.

	FWC-Moderate	FWC-Severe	FWC-Profound
0.25	(0.19-0.26)	(0.12-0.18)	(0.04-0.07)
0.50	(0.3-0.42)	(0.23-0.35)	(0.08-0.14)
0.75	(0.38-0.52)	(0.31-0.48)	(0.12-0.23)
1.00	(0.44-0.6)	(0.37-0.57)	(0.15-0.29)
1.50	(0.54-0.71)	(0.46-0.7)	(0.21-0.39)
2.00	(0.58-0.76)	(0.51-0.75)	(0.26-0.45)
2.50	(0.62-0.78)	(0.54-0.77)	(0.3-0.5)
3.00	(0.65-0.81)	(0.57-0.79)	(0.33-0.54)

Table 4. The floor-wall-ceiling confidence intervals of Jaccard indices at detection radii from ¼ degree visual angle to 3 degrees visual angle around geometry boundaries.

	Obs-Moderate	Obs-Severe	Obs-Profound
0.25	(0.37-0.49)	(0.26-0.3)	(0.04-0.06)
0.50	(0.51-0.66)	(0.43-0.49)	(0.09-0.11)
0.75	(0.58-0.73)	(0.52-0.59)	(0.14-0.18)
1.00	(0.62-0.77)	(0.58-0.65)	(0.19-0.25)
1.50	(0.66-0.82)	(0.65-0.72)	(0.28-0.36)
2.00	(0.69-0.84)	(0.69-0.77)	(0.34-0.44)
2.50	(0.71-0.86)	(0.73-0.81)	(0.38-0.5)
3.00	(0.73-0.87)	(0.75-0.83)	(0.42-0.53)

Table 5 The obstacle confidence intervals of Jaccard indices at detection radii from ¼ degree visual angle to 3 degrees visual angle around geometry boundaries.

	Steps-Moderate	Steps-Severe	Steps-Profound
0.25	(0.33-0.43)	(0.2-0.28)	(0.06-0.07)
0.50	(0.45-0.56)	(0.33-0.43)	(0.13-0.18)
0.75	(0.52-0.62)	(0.4-0.54)	(0.2-0.28)
1.00	(0.58-0.69)	(0.45-0.6)	(0.26-0.36)
1.50	(0.65-0.75)	(0.53-0.67)	(0.32-0.45)
2.00	(0.69-0.79)	(0.58-0.7)	(0.36-0.5)
2.50	(0.73-0.8)	(0.62-0.74)	(0.38-0.54)
3.00	(0.75-0.83)	(0.65-0.77)	(0.4-0.55)

Table 6 The steps confidence intervals of Jaccard indices at detection radii from ¼ degree visual angle to 3 degrees visual angle around geometry boundaries.

Discussion

General Discussion. The experiment conducted here showed a means to derive visibility scores from human tracings of geometry in scenes across simulations of low visual ability, and, further, the work here demonstrated a means to compare those visibility scores to automatically generated visibility scores of scene geometry that are derived from image edges in a scene via the DEVA throughput pipeline. Visibility scores were generated under several different metrics: scaling functions of the distance between an input value, such as a luminance boundary image edge or a participant tracing of an image edge, and a geometric pixel in a task relevant template. The DEVA algorithm's predictive utility was high when compared to participant derived visibility scores; most correlation scores were 0.5 or higher for both metrics in most tasks and across both tasks and acuity level comparisons. The two metrics, Gaussian and reciprocal distance scales, performed nearly equally well in most comparisons and across most scale factors. The average error in the Gaussian metric was higher than in the reciprocal metric making the latter the favorable metric going forward with additional visibility analyses in the future. In addition, the two distance scale factors outperformed derived image metric comparisons; both the root mean square of contrast in scenes and the high spatial frequency content of scenes were not as predictive as simply using the distance to the nearest derived luminance boundary via the DEVA filter. This experiment and analysis therefore shows the utility of deriving luminance boundaries via simulated Low Vision observers, using those luminance boundaries as image edge proxies, and scaling the distance between those image edges and underlying scene geometry edges for predicting the visibility of geometry in public spaces. However, the linear regressions of visibility scores between participants and algorithm show null relationships within acuity levels. This complicates interpretation within single acuity simulations of geometry visibility and the complications are discussed forthwith.

Average participant visibility scores were calculated for each scene in each task across all 3 acuity levels and these average scores were correlated with the DEVA algorithm generated visibility scores for the same scenes and acuity levels. In the first analysis, participant to algorithm comparisons were broken down on a task by task basis across acuity levels with respects to the scaling function used to generate visibility scores; two sets of correlations were produced for both the Gaussian and reciprocal metrics. Across acuities and task by task, the average participant shows a strong similarity to the algorithm visibility score at a scaling factor of 1; however, the algorithm tends to overestimate visibility in each task and overestimation of visibility occurs with both metrics too. The Gaussian metric explains slightly less of the variability in average participant visibility on each of the tasks, and, in fact, a Gaussian scaling factor fails to reach significance in the floor-wall-ceiling task. A linear regression analysis of the floor wall ceiling task with a Gaussian metric indicates a large average error in favor of the algorithm: both a large over prediction of visibility and occasional underestimation of average participant behavior. In truth, linear analyses in each task across acuities indicates large intercepts in favor of the algorithm's visibility values, and this indicates that the DEVA filter is generally outperforming average participant behavior in terms of visibility.

A closer look at the visibility score comparisons for the obstacles task across acuities might reveal why some of the differences in algorithm and participant behavior exist. At moderate acuities, the average participant score is nearly perfectly predicted by the algorithm prediction score: a correlation of 0.917. However, this relationship disappears at severe with r = 0.3440, and, further, the relationship goes in the opposite direction at profound vision loss simulations with an r = -0.662. At moderate levels, participants are clearly outlining the boundaries of objects and the DEVA filter is designed to produce modified canny edge detections of luminance boundaries in the image space. Therefore, if the participants are essentially inferring geometry boundaries from nearby image edges in the image space, participant behavior is likely to match the luminance boundary image provided that both the algorithm and participant are scanning the image for abrupt luminance changes. However, participant behavior drops in similarity dramatically in the severe condition; if the participant is looking for abrupt luminance changes, more and more profound simulations of Low Vision will result in greater difficulty finding the relevant image information useful for inference of underlying geometry. The trend of poor matching with the algorithm's visibility predictions continues into profound simulation losses where the relationship works in the opposite direction, but this is not to say that the data is anti-predictive to human behavior. The correlation here is trending toward 0; there are simply too few data points to reach the appropriate null correlation. Considering the trend towards missing luminance boundaries clearly detected by the DEVA algorithm that are not always being detect by the observer, the participant is likely task switching where the algorithm is not. As scenes become blurrier, participants are loosely interpreting blurry scenes as occasionally not having task relevant geometry despite being told that each trial contains at least one feature of relevant interest: a door, step, obstacle, or door.

In general, in the obstacle tasks and in other tasks as well, there is a large amount of variability in participant performance; participants varied in their ability to do the tasks, their ability to interpret uncertain spaces, and their ability to respond consistently. However, the algorithm always responds to every image with a luminance boundary output, and this highlights the primary difference between the algorithm's predictions and average participant performance. The participant is a task aware observer in that they are uniquely cognizant of the affordances and functional utilities of a public space; they know that parallel lines in a flat monotone area usually indicates a door in a wall. The use of global information then should be a boon to the participant that the DEVA algorithm does not have: an analytical throughput designed to process luminance changes across portions of image space. However, the participant also varies greatly in their willingness to respond at more and more profound vision loss situations. Which is ironic considering that the main reason for the experiment is to probe the visibility of difficult to see geometry in public scenes, but the unwillingness of participants to respond in more profound vision loss simulations goes some length to describe the large intercept in linear regressions weighted towards the algorithm's visibility predictions. The DEVA algorithm's usage of luminance boundaries produces larger visibility scores than average participant performance in the same scenes, but, when one considers the functional differences between algorithm and participant, it is likely that this average increase is exacerbated by peculiarities in human behavior more so than actual differences in detection and usage of luminance boundaries as image edges for geometry inference. In

other words, humans can see blobs and contours at profound vision loss simulations but, unlike the DEVA algorithm, they are somewhat reluctant on average to label those blobs as task relevant geometry.

Notable outliers. When one collapses all tasks together and looks at visibility across tasks and across acuities for either metric, the same trends of participant underreporting yet strong agreement between algorithm visibility and participant visibility can be readily seen with some clear notable differences. Across acuity levels, the ability of the participant to match DEVA algorithm visibility is modestly strong when one uses either the Gaussian or reciprocal scaling function to assess visibility scores. In both cases, the DEVA algorithm generally predicts that scenes will be more visible than average participants are performing. There are some clear notable differences that highlight the difficulties in performing this analysis process.

In scene s2-b at severe Low Vision simulations for the doors task, participants outperform the DEVA filter: 0.5 participant average to 0.3 algorithm performance. The scene has relatively poor contrast, but participants tend to perform well regardless. They perform a task that can be likened to "completing the form" akin to an amodal completion in visual illusions. When a participant sees a long parallel image edge they believe is relevant to inferring a geometry edge with some part of the middle of the image edge being sub detection threshold, it is not unusual for the participant to connect the dots and draw a line over the sub threshold component, and this is likely what happened at severe simulations of s2-b: portions of the door are invisible to the DEVA algorithm.

Another notable outlier is scene 06_17-2 at moderate simulation losses for the floor wall ceilings task where participants' visibility scores are dramatically underwhelming when compared to the algorithm's prediction: an average 0.15 to a 0.68 each respectively. A closer inspection reveals that the expert template differs slightly from a prototypical participant response in that the latter tends to only loosely respond in the correct locations. Participants naturally differ in their ability to infer and trace scene geometry, and some of the tasks and scenes are more difficult than others. This naturally complicates comparisons when the participants find some scenes to be more difficult than the algorithm expects; this scene has very poor scene contrast where the walls connect.

The third prototypical outlier is an image of a lobby with an elevated hallway in the background at severe loss for the floor wall ceilings task where the average participant is dramatically underperforming compared to the algorithm performance: an average 0.10 to a 0.633 respectively. Closer inspection of the expert template and prototypical responses quickly reveals that the discrepancy between algorithm and participant behavior is due to an overly detailed expert template that the participant is reasonably not replicating. The expert template has additional details about obstacles in nooks and crannies in a foreshortened hallway that quickly changes in depth. When the expert templates are generated, they are created in a fashion like the manner in which participant data is scored: scene geometry is scanned for neighboring selections within some distance. However, this scanning process does not account for depth changes in the physical scene. If a scene changes in depth very quickly, objects very far in physical space but nearby in the image space might show up in the expert template such as with tracings that occur in a long hallway.

When one collapses tasks together and only analyzes within single acuity simulations, all similarity between the algorithm and participant behavior disappears entirely. There are no correlations within any acuity level across tasks for either the Gaussian or reciprocal scaling metrics. This is curious given the relatively strong correlations that exist across acuity levels. It appears then that average participant visibility is dropping with Low Vision simulation as expected; performance really is decreasing. The problem is that visibility decreases across acuity are not matching algorithm prediction from luminance boundaries alone. Rate of degradation in visibility is faster than expected for the human observer; they start out with lower visibility performance and worsen in performance faster. If average participant behavior is not predictable at single acuities but is predictable across acuities, it becomes difficult to recommend that the algorithm is accurately predicting visibility in any one scene. However, this is not to say that the DEVA filter throughput lacks predictive power; one simply simulates a broad range of acuities and tests those scenes' visibilities at each acuity level. The analyses comparing average participant to algorithm performances collapsed across tasks and acuities shows that the relative change in visibility as Low Vision simulations intensifies is accurate. Therefore, accurate predictions of the relative visibility of geometry are possible with either distance scaling metric provided one merely simulates a large range of acuities and observes the visibility degradations within.

Comparing scaling metrics. When one looks at the correlations between participant and algorithm visibilities across tasks and acuities, it would appear then that either the Gaussian or reciprocal metric would be a nearly identical choice, but a closer inspection of the linear analyses for both metrics across all tasks and acuities reveals a different story. It is possible to directly compare the relative performances of metrics by comparing the residual errors in their linear regressions across relevant conditions. In this case, linear regressions were performed for both metrics separately across all tasks and acuities; this produces nearly identical regression coefficients between the two metrics. However, one can regress the residual errors of the two metrics' linear regressions to reveal which metric has a larger pattern of average error. If both metrics had the same amount of error in predicting participant visibility scores, the regression of residuals would produce a slope of one, yet the slope of the regression is weighted towards the Gaussian metric. By taking the standard deviation of the residuals in a task by task basis for both metrics, one can see that the average standard deviation between residuals in individual tasks differs between metrics as well. Therefore, the reciprocal metric at a scaling factor of 1 outperforms the Gaussian metric in terms of predicting human visibility scores. However, a word of caution is necessary here as the difference between the two metrics is almost arbitrary due to the free parameter used to scale the distance across the image space. The reciprocal metric appears to outperform the Gaussian metric at a scale factor of 1, but this scaling factor was chosen arbitrarily. A correlation of visibility scores between the two metrics reveals near perfect correlations; the differences between the two metrics are very slight and subtle. Therefore, the choice of metric is nearly arbitrary and evidence here appears to indicate the one could fit either metric to human performance data just by adjusting the free parameter appropriately

Image processing discussion. The second part of the analysis process probed the nature of image processing and how much visibility could be predicted comparing visibility

scores to derived image properties across acuity level simulations. The DEVA filter throughput is predicated on the notion that simulating Low Vision observers and luminance boundaries would be superior to any prediction that would arise from analyzing images alone. DEVA simulates Low Vision by utilizing a compressed CSF to convolve images such that they lose both spatial frequency and contrast information equally. If visibility behavior can be predicted adequately at rates better than either of the distance metrics, it would obviate the DEVA algorithm process; one would need only perform discrete Fourier transforms to analyze image properties and subsequently predict scene visibility. Therefore, additional analyses of scene properties were performed to understand if DEVA can be outperformed by analyzing the statistics of image properties.

A high spatial frequency (HSF) analysis was performed that detected the highest spatial frequency component of a scene at the lightest available contrast available in that scene, and it was found that HSF typically does not correlate with any of the average participant performances except for obstacle detection when one considers data by task and across acuities for both distance metrics. Steps, doors, and floor wall ceiling tasks failed to reach significance for any of their linear analyses, and it is likely because the participant as a functional observer could utilize global information to perform those tasks. Observers do not need fine local information to detect that large parallel lines in space that typically indicate the presence of doors, and, likewise, the same can be said of both steps and floor wall ceiling tasks: things that tend to be very large. However, obstacles in the walking path tend to be small such as chairs and tables, but it is not always the case that a chair or table is small in the image plane. Obstacles do tend to be smaller than entire staircases though, yet the variety in size of obstacles in scenes may go some length to explain why high spatial frequency was not very predictive of visibility scores even in this one task.

One might ask why one would bother performing an analysis of HSF to visibility score given that HSF should be highly correlated with any visibility related metric; blurrier images should be harder to see a-priori. Observer figure 21. As Low Vision simulation intensities increase, the average HSF in cycles per degree decreases in each set of images for all tasks, but the variance of HSF within those image sets remains large regardless of the average decrease in HSF. It is tempting to think that this variation is due to image artifacts that create sharp bands in profound Low Vision image simulations.

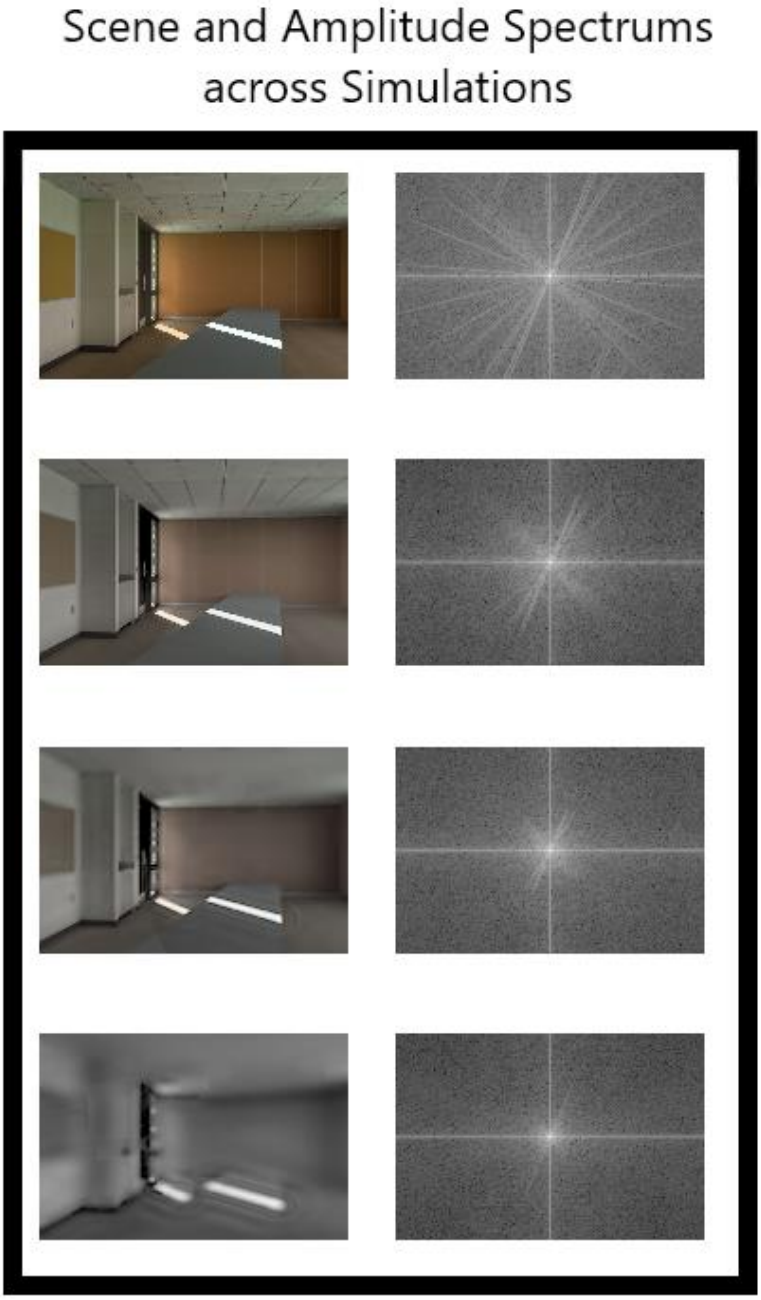


Figure 24. A set of images and progressive Low Vision simulations from moderate to profound intensities. The amplitude spectra of the discrete fast Fourier transforms are plotted next to their respective images. As Low Vision intensity progresses, high spatial frequency components are attenuated, but, in the profound Low Vision simulation, some HSF components remain. This scene is one of several that has strong artifacting in progressive Low Vision simulations. These can be seen around the light band on the floor.

As can be seen in figure 24, high spatial frequency components remain at even very profound Low Vision simulations: a situation where almost none should exist. It is then possible that the large variances in HSF in each set of images could be due to artifacting resulting from the prototype image processing throughput. However, it could also be the case that a large amount of variance high spatial frequency in cycles per degree simply exists from scene to scene. Observe figure 25. The HSF content of all scenes was scanned to find images with similar peaks in cycles per degree visual angle. Two groups were formed: one of images near 8 cycles per degree and another near 4. The former is a moderate Low vision simulation of several different scenes and the latter is the same but at profound Low Vision. Each image is from a different task, in the order of doors, floor-wall-ceilings, obstacles, and steps. Despite the apparent variation in scenes, the peak HSF for all scenes within their respective group is highly similar. It then seems that HSF naturally varies a great deal from scene to scene under normal conditions. Some of the images plotted in the profound grouping here contain artifacts as well. Not all of them do and they all have similar peak HSF information regardless. Given that the HSF varies widely from scene to scene within acuity levels and that this variance does not seem dependent on image artifacting, it is possible then that, despite expectations, HSF potentially does not directly correlate with participant derived visibility scores.

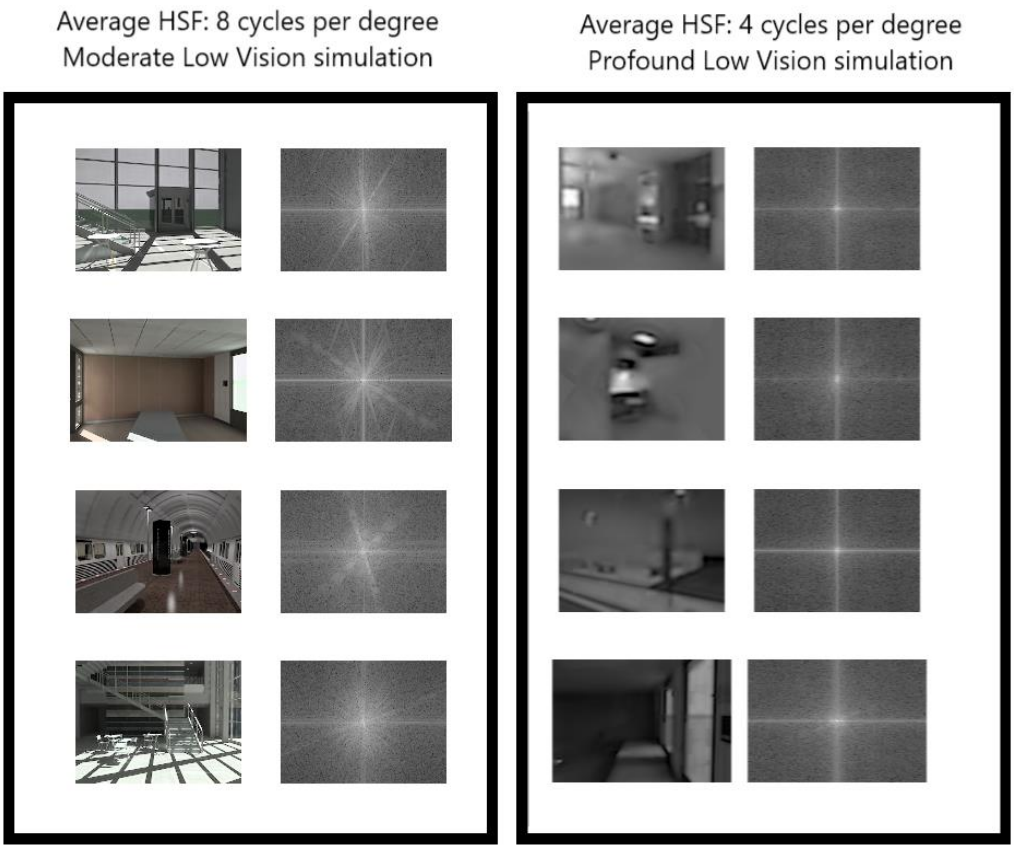


Figure 25 Two sets of images and their respective amplitude spectra. The left group of images shows moderate Low Vision simulations with HSF information at around 8 cycles per visual degree angle. The right group contains images that shows profound Low Vision simulations with HSF information at around 4 cycles per visual degree angle. Each image in either set is very different from one another: some scenes are sparsely decorated while others have many small details, doors, and chairs. Despite this, the HSF for each of the scenes within either group are like their neighbors.

An analysis of the scene contrast was performed that did not detect any meaningful connection between changes in the RMS contrast across scenes and average participant derived visibility scores. This is particularly surprising considering the prototype algorithm's image manipulation process: a processing throughput that simulates Low Vision via heavy reductions in spatial frequency, color saturation, and contrast. However, a closer inspection of the images themselves reveals a likely culprit that could potentially spoil an RMS based analysis. The initial prototype algorithm occasionally produced heavy image artifacts around strongly illuminated surfaces as can be seen in figure 23. Scenes with visible light bulbs, overhead lighting, beams of light on the ground, and surface highlights were susceptible to image artifacts generated by the prototype algorithm's image processing. The prototype algorithm is under current and frequent development. New developments in the way the prototype algorithm handles

Low Vision image simulations allows the production of Low Vision simulations with less image artifacts.

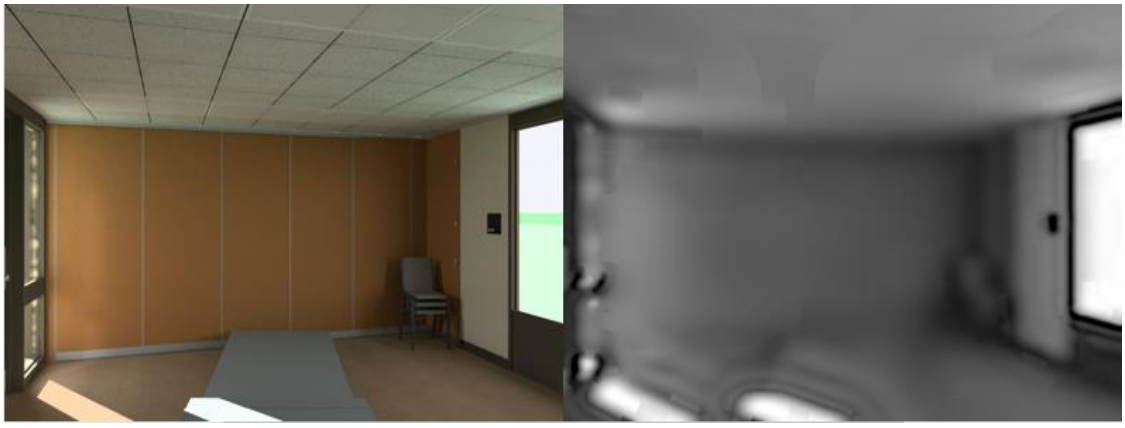


Figure 26. An example of intense image artifacting that occurs in the version of the prototype algorithm used in this experiment. The light band on the floor has strong contrast from rings that are introduced by image processing in the DEVA algorithm's simulation of Low Vision. More recent versions of the prototype have much less pronounced image artifacts, but participants did not view stimuli generated with those later versions.

It is tempting then to redo the RMS analysis with newly generated HDR scenes from an updated prototype algorithm, but a word of caution is due before undertaking that action. The experiment undertaken here was done on an active in-development prototype, and, as such, participants were tested on images that contained image artifacts. A comparison of prior generated visibility scores to the RMS of newly generated scenes might be suspect in that the latter could have a confounded relation with the former. Participants' visibility scores were not generated specifically for those artifactless images but instead were generated via scenes that contained artifacts and thus have different luminance boundaries. A central assumption of this research was that the observer infers geometry boundaries from luminance boundaries. The prototype algorithm calculates visibility scores from luminance boundaries in the simulated Low Vision image, and, if luminance boundaries change because of a maturation in the prototype algorithm, a new experiment must be completed before comparing visibility scores between participants and the prototype algorithm. Participants were affected by the artifacts in the more profound Low Vision simulations, and they would likely be affected by the absence of artifacts as well. Any comparison between prototype algorithm and participant behavior therefore requires a solid foundation of identical images between the two. Because of the image artifacts in the current stimuli set and the RMS contrast metric being susceptible to such artifacts, it is difficult to draw any predictive conclusions between contrast and geometry boundary visibility.

Stepwise Jaccard Analysis Discussion. The stepwise Jaccard analysis shows growing agreement between the prototype algorithm and observer tracings as detection radii are expanded outward from geometry boundaries. This trend remains the same in both

moderate and severe vision simulations. However, there is a significant difference in the profound vision loss condition between the prototype algorithm's luminance boundaries and observer tracings; the observer is typically under responding to most scenes. One can see this in the largest detection radius where even at three visual degree radii around all geometry boundary pixels in the expert templates, agreement between participant and algorithm remains very low. The 95% confidence intervals of Jaccard indices at each stepwise distance do not overlap with any of the other conditions. Even with a casual look at the data, it is easy to tell that participants were far superior at finding door geometry in scenes. If one were to perform the Jaccard analysis again with this superimposed image of 10 different participants, there would be higher agreement with the prototype algorithm's detection of geometry boundaries. It is posited then that observers are generally under-reporting information in the scene, and that the prototype algorithm's tendency to overestimate visibility is likely more due to participant behavioral tendencies than actual visibility of geometry boundaries.

Conclusions. It appears then that the prototype algorithm throughput is strongly predictive of human visibility and that the prototype algorithm outperforms derived image statistics. However, the prototype's use of luminance boundaries in images is currently predicting greater visibility than is found in average participants at any single acuity level. Average participant visibility scores are predictive across acuity simulations but not within them. If one wants to predict the visibility of a public space with this analysis process, it would be wise then to simulate multiple acuity levels and create visibility predictions across them; the relative change in visibility through the spectrum of acuity levels is relatively accurate. Furthermore, a step-wise Jaccard analysis showed strong agreement in distance-based Jaccard indexes of geometry boundary visibility shared between human observers and the prototype. Further analysis showed that human observers are highly likely to be underreporting in more profound vision loss situations and that this likely has led to an arbitrary overestimation of visibility. Furthermore, the deva-filter subroutine of the algorithm prototype works by convolving images with parameters derived from a simulated Low Vision observer's hypothetical CSF: attenuations in high spatial frequency components and contrast sensitivity alike, but, even so, the prototype's process of deriving nearest neighboring luminance boundaries manages to outperform both kinds of derived image statistics that are central to the simulation of Low Vision.

Therefore, the prototype algorithm's throughput processes are a superior method of predicting visibility in public spaces than merely analyzing local scene statistics. It is likely that observers are scanning the image for abrupt changes in luminance akin to the DEVA algorithm's process, but questions remain about differences in strategy between participants in this experiment and the DEVA algorithm's luminance boundary strategy. Observers appear to be gathering information across the image space necessary for their tracings; if this were not the case, there would be no correlation between high spatial frequency content in scenes and tracing behavior in obstacle detection tasks. Interestingly, participant behavior changes during obstacle detection as acuity simulations worsen, yet the algorithm's performance does not degrade at nearly the pace of the participants across acuity. The question is then begged about the prototype's assumption

of luminance boundary information; if this information is so critical to the observer, why is the filter predicting visibility much higher than average participant performance?

It is possible that peculiarities in participant response have resulted in this difference, but it could also be the case that observers in simulated Low Vision tasks are not sensitive to information in the scene necessary to their decisions. In other words, the participants in the task might not be as sensitive to changes in luminance in the same way as the prototype algorithm or that they are unable to use this information because of their predilection towards responding in a task functional manner. It is therefore necessary to ask how much spatial information is needed to understand the scene in a functional way that matches participant behavior; how much of task relevant image does an observer need to see to understand an obstacle's identity.

Chapter 4. Identifying Objects Inside Scenes with simulated Low Vision.

Imagine asking someone to identify every animal in a zoo while only being able to view the world through a small aperture. During the task, it's likely one would spend most of their time scanning each prospective animal over and over; the observer would saccade over the animal repeatedly to build a Gestalt of the animal's form and thus identify it. Some animals might be easier than others because of unique and highly salient features such as tigers, but others, such as small marsupials, might be much more difficult to identify due to their lack of clear identifying features. Either way, in each case, the observer's behavior is the same: scan the image space until sufficiently saturated by task relevant information.

Currently, the work in DEVA suggests that visibility of geometric boundaries in a public space is modestly predicted by the distance if geometry boundaries to their nearest luminance boundaries in the image space of the scene. Derived image properties such as the high spatial frequency of an image also correlate with visibility of underlying geometry. However, when participants are looking at scenes and tracing the objects therein, a bevy of global scene information is available to the participant. In the previous experiment, participants would often trace over portions of geometry boundary pixels that were predicted to be invisible by the prototype algorithm. The tracing of apparently subthreshold geometry boundaries frequently occurred in situations where a subthreshold geometry boundary was surrounded by predicted-to-be visible geometry boundaries. Thus, participants appear to occasionally "connect the dots" when looking for scene geometry in simulations of Low Vision.

Given that observers can infer invisible geometry boundaries when those boundaries are nearby neighboring visible geometry, it seems likely then that participants are gathering spatial information across the image space necessary to understand the geometric segmentations of scene objects and to trace the contours of those objects within the scene. This phenomenon is especially noticeable in Low Vision simulations of public scenes where objects, such as stairs or doors, have a predictable geometric regularity, but this regularity is only apparent at a global scale as any singular view of local stair corners quickly becomes impossible to interpret at progressive levels of Low Vision. It is likely that the observer is interpreting the global nature of the scene by saturating themselves on relevant local image features congruent to the concept of some known geometric construct: stairs have corners, doors have parallel lines, atriums have empty foreshortened boxes.

It has long been known that images contain local image statistics that are potentially exploitable by the observer, but the extent of the observer's use of this information for scene understanding is not completely understood. Research by McDermott (2004) tested whether human observers could successfully identify junctions of luminance boundaries, such as T or Y intersections, in progressively larger and larger image patches containing those image edge junctions. When image edge junctions were revealed in image patches that had a circumference of 13 pixels, observers performed at

chance levels in discriminating whether a junction was present in the patches presented. However, as image patches grew in diameter, observer's ability to discriminate the presence of junctions increased; at 200-pixel diameter image patches, observers had an average accuracy of patch classification of 90%. To understand whether an image patch contains an edge junction requires some understanding of the 3D structure of the scene or some expectation of the scene's structure; the local contrast values of some image edge are not a priori meaningful to understanding geometry edges for an observer without some understanding of the global context of the image. This was further investigated by DiMattina et al (2012) who looked at the differences between image patches that contained hand-labeled occlusion edges and image patches that were labeled as within-surface boundaries by observers. DiMattina et al compared human observers' ability to understand causes of edges in images to automatic linear classifiers and neural network analyses predictions of those image edge causes. Observers, both human and algorithm, improved in inference of edge discrimination with increasing patch size, and both the human observers and neural network classifiers improved with scale of image patch presented. DiMattina et al also showed that neural network classifiers require scale information and some knowledge about the image edges location in the whole image to be able to perform comparably to human observers. The fact that an algorithmic approach requires coordinate space knowledge of luminance boundary location to match human performance implies that human observers are aware of the relative positions of most task relevant luminance boundaries in the image space and that the human observer is reconstructing that coordinate space presumably for geometry boundary inference. This notion of tracking many hundreds of coordinates of luminance boundaries is simpler than it sounds and tracking can be performed with simple heuristics that interpolate between coordinate areas. For example, it is understood by all observers that parallel lines in doors, stairs, chairs, and so on usually continue even after apparently disappearing behind an obstacle in the image.

As observers scan a whole image of a public scene, they are implicitly compelled to understand the 3D structure of a scene by the task of imagining navigation through it. To understand the geometry makeup of some scene requires the observer to gather luminance boundaries from the image space and make inferences about the geometry boundaries that these luminance boundaries possibly represent. If the observer's goal is to navigate around 3D space and the obstacles therein, it behooves the observer then to be able to accurately infer the cause of luminance boundaries in the image space available to them else they run the risk of making a poor inference and potentially running afoul of some unexpected obstacle. Work on natural scene statistics (Vilankar et al, 2014) suggests that local statistical information about the cause of luminance boundaries exists for the observer to exploit. The researchers asked observers to classify edges in images as either occlusion or non-occlusion edges and then further subclassified non-occlusion edges as either reflectance, surface changes, cast shadows. By taking observer classifications and comparing them to local contrast statistics for each of the categories, Vilankar et al found that there were average differences between each of the categories in terms of their average contrast values, but it is not clear if these average differences are

utilized by observers to make inferences about 3D space rather than merely being coincidental with their categorization of edge cause. Furthermore, it is also unknown how the observer could potentially gather and summate the potential information from these local contrast differences to build an understanding of the global structure of an image.

The problem of the observer potentially utilizing local image statistics to infer causes of edges and thus begin to infer 3D scene structure is further compounded by a Low Vision observer; to what extent does Low Vision degrade the observer's ability to detect and infer geometry boundaries in an image that has lost information? One way to examine this question is to probe the predictive qualities of images; if local statistical information exists to be exploited by the observer, it is presumably possible to predict the accuracy with which some observer might be able to categorize scenes by knowing statistical properties inherent to those scenes. For example, if one were to ask observers to classify scenes into object categories, e.g. whether they contained certain classes of obstacles, it might be possible to predict scene classification accuracy by knowing certain image statistics or pre-derived image qualities. The root-mean-square contrast is an image metric that describes overall contrast in complex scenes and it might have predictive value for anticipating scene classification accuracy. Another possible metric is the high spatial frequency cut-off of each scene; the high spatial frequency information available to observers might predict the rate of scene classification accuracy. Finally, the previous experiment derived scene visibility scores via by scaling distances between geometry boundary pixels in scenes and Low Vision derived luminance boundaries in Low Vision simulation images. It could be the case that these derived visibility scores are predictive of scene classification accuracy. Furthermore, it could be the case that any of those three metrics are predictive of scene classification accuracy, but which one and to which extent is unknown.

A three-alternative forced choice task is proposed where observers will see images of public scenes and must identify if those images contain either a door, a chair, or neither a door or chair. In addition, several simulated Low Vision conditions are proposed: moderate, severe, and profound vision loss simulation intensities. If observers are dependent on some aspect inherent to the scene, scene classification accuracy should be predictable from the relevant aspect. It is likely that image statistics such as the highest spatial frequency or the root mean square contrast of scenes are predictive of classification accuracy, but it is also possible that distances between geometry boundaries to luminance boundaries in the Low Vision simulated images are predictive of classification accuracy as well. Image statistics and distance measures are explored with the intent of predicting the rate at which observers can classify scenes while viewing them under simulated Low Vision. In the previous experiment, scaled visibility scores were predictive across but not within acuity levels of obstacle tracing accuracy; so, it is predicted here that classification accuracy can be predicted ahead of time if one knows the relative degradation of images in a stimulus set in terms of their RMS, HSF, or visibility scores and the spectrum of Low Vision simulation. In other words, one can likely predict average classification accuracy at any one Low Vision simulation in a task

like this if one knows the spectrum of Low Vision used and the average change in metrics across stimuli.

Methods

Stimuli. Scenes were high dynamic range (HDR) images generated in Radiance 3D Rendering (Ward & Shakespeare, 1998). The images were physically realistic renderings in terms of lighting and reflectance of surfaces. The images were displayed on a MultiSyinc E243WMi LCD monitor with a maximum brightness of 200 cd/m^2, a resolution of 1600 x 1200, and a refresh rate of 120 Hz. HDR images have a much larger range of luminance values than can be displayed on this current experiment's monitor, and to ensure accurate display of the HDR images with respect to keeping local contrast values linear across the image while also displaying the HDR image in a manner that is physically possible for the current monitor. The images were displayed in RGB with 256 color-mapping in a sRGB color space. The monitor was gamma calibrated such that the luminance values from the monitor were linear. The tonemapping was linear up intil luminance values were twelve times the median luminance for the scene, after which the values were capped a maximum luminance value. When displayed, the images had an average luminance of 102.7 cd/m^2. The images were displayed with a resolution of 1200 by 1400 pixels.

The HDR images were physically realistic representations of a real public space: an atrium in the University of Indiana. For architectural design evaluation, each image is generated to represent the field of view and perspective of discrete observers. The images used here were not originally designed for experimental manipulation. The field of view for each simulated viewpoint varies from image to image; the visual angle across the simulated observer's field of view both horizontally and vertically changes from scene to scene. To ensure that participants in this experiment experienced the appropriate visual angle in each scene, it would have been necessary to have moved participants back and forth in the middle of the experiment because of the randomized presentation of stimuli. The stimuli had field of views that ranged from 70 to 50 visual degree spans horizontally and 50 to 45 visual degree spans vertically. Participants were seated at 30-cms from the screen. This viewing distance provided a 35 visual-degree vertical span for each stimulus. This distance was chosen to provide the largest visual angle presentation of stimuli without seating participants so close to the screen that they could count individual phosphors and while still being at least a factor of 2 within the largest vertical FOV span.

There were 12 HDR images in total, and the images were split into 3 different scene categories: door images, chair images, or images that had neither doors or chairs. The images were simulated at 3 different levels of Low Vision using the DEVA prototype algorithm (Thompson et al, 2017; https://github.com/visual-accessibility/deva-filter). In total, there were 36 HDR scenes across simulated acuity levels, and each scene was presented ten times during all three of the simulated acuities. Throughout the experiment, a participant saw 360 HDR scenes.

Participants. 3 graduate students from the University of Minnesota took part in the experiment. The experiment was approved by the Institutional Review Board of the University of Minnesota. Each participant completed a single experimental session that

lasted about one hour, and each participant completed all three of the image manipulations during their participation. The participants all had normal or corrected to normal vision and they were not compensated for their participation.

HSF analysis. High spatial frequency cut-offs of the power spectrum of scenes are calculated as part of a multi-step Fourier analysis process. First, the amplitude spectrum of a scene is found. This is done by taking the Fourier transform of an image which produces the amplitude spectrum of the scene with complex numbers. The absolute value of the squared values of that analysis produces the power spectrum of the scene. The second step is to take the power spectrum and zero-point the graph by shifting zero frequency elements to the center of the power spectrum. The goal of this analysis is to summarize the cut-off in the amplitudes across all spatial frequency orientations in the image space, and, to do that, it is necessary to average across orientations in the power spectrum. Therefore, the third step is to interpolate an averaging function across the spectrum in polar coordinates. This produces the average contrast power of spatial frequencies across the image space. The fourth step is to calculate a cut-off frequency for some criterion level; the choice of criterion is arbitrary and here the assumption is that the visual observer cannot be sensitive to some contrast power below 0.001. The highest spatial frequency cut-off in a scene is then solved for by solving a linear regression for the interpolation function and returning the expected frequency for the arbitrary contrast criterion.

RMS analysis. Contrast reduction is the second part of the prototype algorithm's Low Vision simulation process. Contrast is the difference in luminance between a pixel and a reference luminance. The contrast of a sinewave grating is typically determined with a simple contrast function such as Michelson's contrast: the difference in luminance between brightest and darkest spots divided by the sum of luminances from those same spots. However, simple measures of contrast are not meaningful for entire complex images. To measure the relative contrast in HDR scenes, one needs a measure that spans the entire image space and gives contrast as a function of the average deviation in luminance across that space. For that purpose, one can employ a definition of contrast known as the root-mean-square of contrast luminance levels. RMS contrast is the standard deviation of pixel luminance in an image divided by the average luminance of the entire image. For analysis purposes, RMS contrast for each scene in the experiment is compared to the average participant generated visibility for those scenes.

Expert Geometry Templates. Prior to analysis, it is necessary to create subsets of geometry boundaries that function as baselines for accurate geometry boundary detection; these subsets of geometry boundaries are known as expert templates. Each expert geometry template is unique to the scene and task from which they are created and each one contains all relevant geometry boundaries for some scene at some task. As seen in figure 10, many images of public scenes contain a great deal of geometry boundaries, but only a small subset of those boundaries are relevant for an ambulating pedestrian at any given point in time.

Visibility. Visibility scores are produced in a multi-step process that begins by taking the relevant expert template for a scene, and then, for each geometry boundary pixel in the expert template for that scene, measuring the Euclidean distance between a geometry

boundary's pixel location and the nearest input pixel, either a luminance boundary pixel from the prototype process or a tracing pixel from the participant, and, finally, scaling that distance with some visibility metric The utilization of subsets of geometry boundaries as expert templates allows one to enforce task specific analyses of visibility. Once every geometry boundary pixel has an associated Euclidean distance measure, it is necessary to interpret those distance scores.

In the final stage, visibility scores are created by interpreting the distance score between each geometry boundary pixel and their nearest neighbor input. At this early stage of validation, any chosen metric need only produce lower visibility scores when the geometry and image edge pixels are far apart and high scores when they are closer together. Two simple distance metrics are employed here to create visibility measures as a function of distance: a reciprocal scaling metric:

$$\text{Visibility} = \text{scale} / (\text{degrees per pixel} * \text{input} + \text{scale})$$

and a Gaussian scaling metric

$$\text{Visibility} = e^{\wedge} (-1 * (\text{degrees per pixel} * \text{input} / \text{scale}) ^{\wedge}2)$$

where input is the Euclidean distance between geometry boundary pixels and input pixels, scale is a free parameter, and degrees per pixel is a conversion factor that turns the Euclidean distance between pixel locations into visual degrees. This conversion factor varies very slightly from scene to scene and is determined by the field of view of the simulated observer when the CAD model specifies a viewpoint to be rendered into an HDR image. The reciprocal scaling metric is one that creates visibility scores that steadily drop to 0 as the nearest image edge input moves away from the relevant geometry pixel, and the Gaussian scaling metric produces visibility scores that remain strong at interim distances before quickly advancing to 0 with greater and greater distances. Both metrics have a free parameter that can either be set to any vale which allows for an adjustment of fit between algorithm visibility scores and human participant visibility scores.

Procedure. Prior to the experiment, participants were briefed on the demands of the task, the nature of Low Vision and people with low visual ability, and that they would be using the keyboard to identify the contents of images. Participants were told that they would be shown Low Vision simulated images of public spaces, and the participant was informed that they must identify the contents of these images. They were told that the images would be shown to them briefly and that they must make a guess after the image disappeared as to what it contained. Participants had a forced choice of three different categories: chairs, doors, or neither chairs or doors. Stimuli were presented in moderate, severe, and profound acuity blocks, and the doors, chairs, and neither-doors-or-chairs scenes were presented in equal numbers randomly throughout the blocks. Images were shown for 500-ms before an answer was allowed from the participant. Upon pressing a key, the next stimuli would appear, and the process began anew until all 120 stimuli for the acuity block finished. To reduced learning effects commonly seen in blurry stimuli sets, participants began on the profound vision loss simulation and moved towards

moderate Low Vision simulation intensity. After the participant finished all 360 trials, they were debriefed on the experiment.

Results

Predictive visibility scores generated for each individual scene were regressed against average participant accuracy for each of those scenes to test if one could predict classification accuracy on stimuli by stimuli basis using visibility score alone, and it was found that this was not the case ($F(1,34) = 0.43$, $p=0.516$, $R^2= 0.0125$). Average visibility scores were generated across images within individual acuity blocks, and averages across all three stimuli sets were acquired. These averages were regressed against each of the participants' average classification performance per acuity level. It was found that the average visibility scores across sets of stimuli were highly predictive of scene classification accuracy ($F(1,7) = 62.7$, $p<0.001$, $R^2=0.9$)

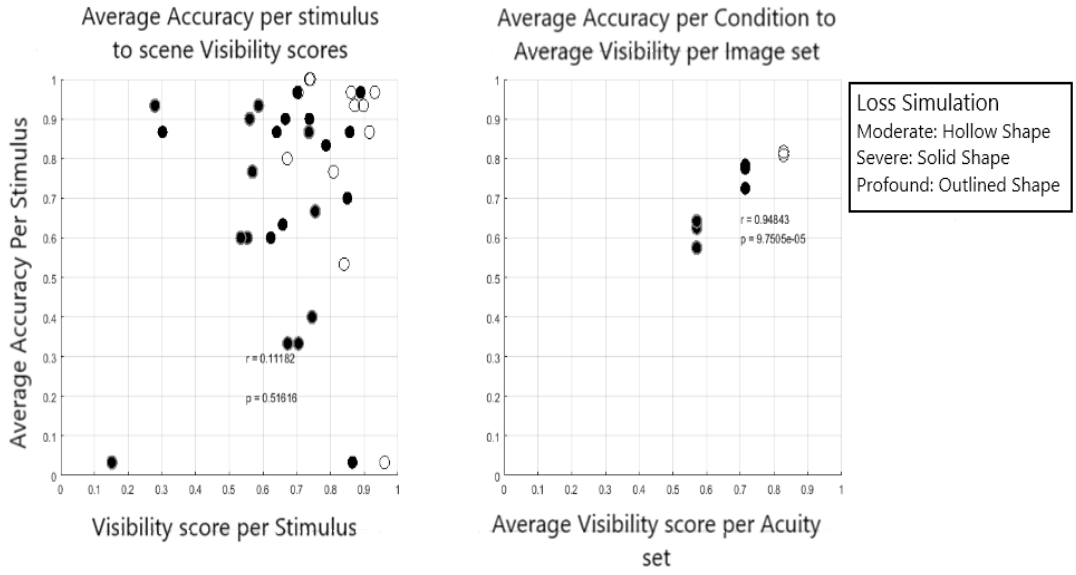


Figure 27. The left graph is the linear regression of average accuracy per scene compared to the visibility score of that scene across acuity levels. The graph on the right is the average accuracy per acuity block per participant to the average visibility score across images in a given block repeated across acuity levels. The visibility score of any one scene is not predictive of classification accuracy, but average classification accuracy was highly predicted by knowing the average visibility scores in image sets across acuity levels.

The RMS contrast of each scene was regressed against the average participant accuracy for classifying those specific scenes individually to test and see if the former was predictive of the latter, and it was found that RMS contrast was not predictive of average classification accuracy for individual scenes ($F(1,34)=0.15$, $p=0.701$, $R^2= 0.00439$). The average RMS contrast was calculated for the stimuli set of each acuity

block, and these average RMS contrasts were regressed against the average classification performance of individual participants in those blocks. The average RMS contrast across stimuli sets was found to be highly predictive of scene classification accuracy ($F(1,7)=37.4$, $p<0.001$, $R^2=0.842$)

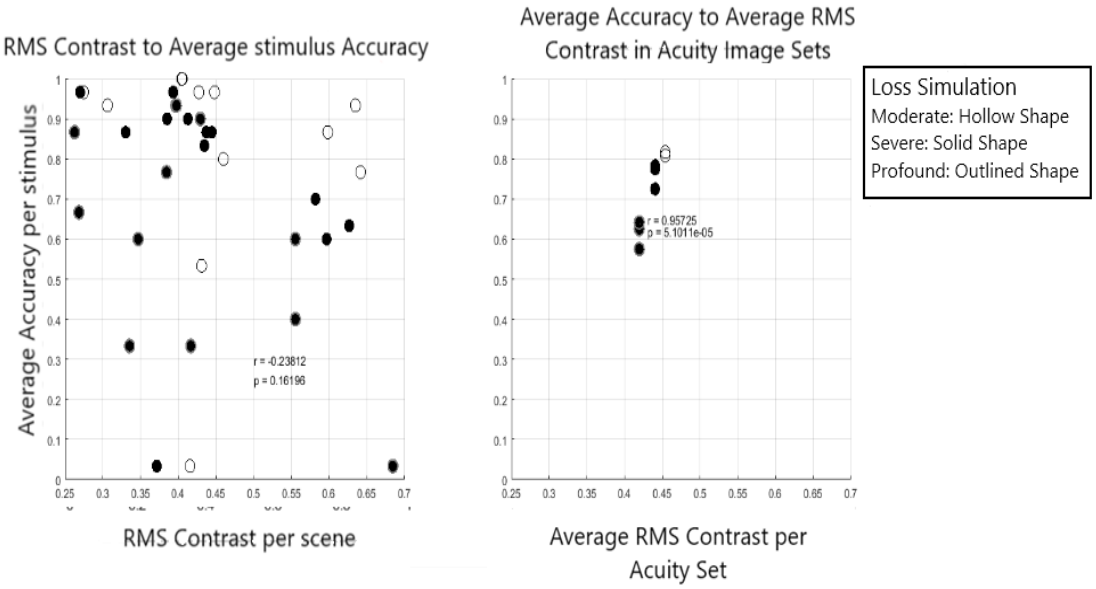


Figure 28. The left graph is the linear regression of average accuracy per scene compared to the RMS score of that scene across acuity levels. The graph on the right is the average accuracy per acuity block per participant to the average RMS score across images in a given block repeated across acuity levels. The RMS score of any one scene is not predictive of classification accuracy, but average classification accuracy was highly predicted by knowing the average RMS scores in image sets across acuity levels.

The high spatial frequency cut-off per visual angle of each scene was calculated and regressed against the average classification accuracy for each scene to see if HSF within individual scenes was predictive of scene classification accuracy, and it was found that this was not the case ($F(1,34)=2.04$, $p=0.162$, $R^2= 0.0567$). The average HSF within each set of acuity block stimuli was calculated and regressed against the average classification performance of participants in those acuity blocks, and it was found that the average HSF across stimuli sets was strongly predictive of average scene classification ($F(1,7)=76.7$, $p<0.001$, $R^2=0.916$)

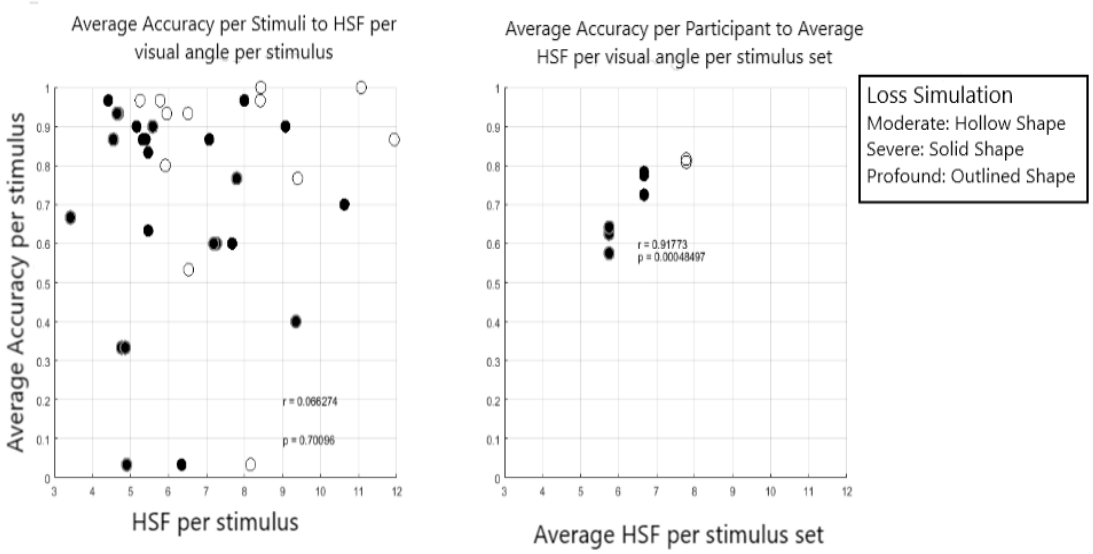


Figure 29. The left graph is the linear regression of average accuracy per scene compared to the HSF score of that scene across acuity levels. The graph on the right is the average accuracy per acuity block per participant to the average HSF score across images in a given block repeated across acuity levels. The JSF score of any one scene is not predictive of classification accuracy, but average classification accuracy was highly predicted by knowing the average HSF scores in image sets across acuity levels.

Discussion

Linear analyses found that average classification accuracy for individual stimuli was not predicted by RMS contrast, HSF cut off, or geometry boundary visibility scores derived for those scenes. However, regression analyses did find that one could very strongly predict average scene classification accuracy within acuity blocks by knowing the average value of any of the three metrics across the span of the simulated Low Vision spectrum used in the experiment. In the previous experiment, within acuity analyses between predicted visibility and actual visibility of geometry boundaries were disrupted by large differences within sets of stimuli. This can be seen again in figures 21 to 23; there is a very large amount of variation in all three sets of images in terms of both their metrics and average classification accuracy. The act of averaging within acuity blocks removes that variation and provides predictive value when one has multiple averages of metrics across a spectrum of simulated Low Vision. Therefore, it is possible to predict scene classification accuracy within image sets provided one has simulated that image set across multiple Low Vision intensities and measured the average metrics of images within those image sets. Even more curious is that it is easy to see that all three of the metrics used here are highly correlated with one another, and all three give similarly very strong predictions of average classification performance in single acuity blocks within a simulation spectrum. This complicates the choice of picking a superior predictive metric,

but their correlation was likely, in hindsight. RMS contrast and HSF cut off are both intrinsic parts of Low Vision simulations and the distance of luminance boundaries in the Low Vision simulation to underlying geometry boundaries is a tertiary function of the degradation in the image space caused by reductions in contrast and spatial frequency. The choice between the three is almost redundant if one wants to predict how accurate observers will be in classifying scene accuracy under Low Vision simulations in general.

During the task, observers ran in three acuity blocks where the most intense Low Vision simulation was first followed by the second least intense and so forth. This was done purposefully to prevent the participant from learning the stimulus set. In previous pilots and, even in previous pilots of experiments in this thesis, participants often had higher than expected accuracy when they were intimately familiar with the scene. Low Vision individuals often recognize furniture in their own home, at work, or in public spaces due in part to the intimate knowledge they have of those locations. In addition, stimuli were presented at 500 ms to prevent a participant from staring at and memorizing individual images. Research suggests that accurate scene classification can occur at presentation times as low as 14-ms, and, with 500 ms, participants had ample time to understand and respond to stimuli (Potter, 2012). The timing delay and presentation order were consistent across all stimuli in all blocks and with all participants. If one were to test from moderate to profound intensities with no time delay, the experiment would essentially be partially due to scene learning rather than instant recognition of scene contents.

In previous DEVA related experiments (Chapter 3), it was found that one could predict the visibility of underlying geometry boundaries by simulating Low Vision, taking a canny edge detection of that simulation, and then measuring and scaling nearest neighbor distances between the two. That research is expanded on to show that the same distance metric can be used to predict average scene classification accuracy. The previous experiment suffered from large variation in the stimulus set: an ecological factor representative of the real world. With a limited stimulus set, the current experiment suffered from the same foible. It seems then that both experiments are circling the real problem; there are intrinsic factors to images simulated at Low Vision and observer behavior that allow prediction of scene visibility. If this were not the case, deriving average metrics in image sets across spectrums of Low Vision would lack predictive power. Steps must be taken then to understand the capacity with which observers are able to gather image information and use that information for scene understanding.

The prototype algorithm is designed with the assumption that luminance boundaries and distances from them are inherently meaningful and useful for observers wanting to derive understanding of underlying geometry from the scene image. This assumption is valid both in cases where the observer must classify scenes in general or trace their geometry boundaries exactly. In the former, scaled distances between luminance boundaries and task relevant geometry boundaries predicts average classification accuracy of scenes so long as one has both derived metrics for all images in

a set and done so across multiple Low Vision simulation image sets. Like in the previous experiment, prediction of a simulated Low Vision observer in this task rests on the nature of the spectrum of Low Vision that has been simulated. Previously, it was the case that the observer's behavior could only be predicted across acuities and trend holds in this experiment as well. Effective prediction of Low Vision simulated observer behavior rests on the ability to simulate across spectrums and derive meaningful metrics for prediction.

Bibliography

Bailey, I. L., & Lovie, J. E. (1976). New design principles for visual acuity charts. American Journal of Optometry and Physiological Optics, 53(11), 740–745. http://doi.org/10.1097/00006324-197611000-00006

Bochsler, T. M., Legge, G. E., Gage, R., & Kallie, C. S. (2013). Recognition of ramps and steps by people with Low Vision. Investigative Ophthalmology and Visual Science, 54(1), 288–294. http://doi.org/10.1167/iovs.12-10461

Bochsler, T. M., Legge, G. E., Kallie, C. S., & Gage, R. (2012). Seeing steps and ramps with simulated low acuity: impact of texture and locomotion. Optometry and Vision Science : Official Publication of the American Academy of Optometry, 89(9), E1299-307. http://doi.org/10.1097/OPX.0b013e318264f2bd

Boucart, M., Despretz, P., Hladiuk, K., & Desmettre, T. (2008). Does context or color improve object recognition in patients with Low Vision? Visual Neuroscience, 25(5–6), 685–691. http://doi.org/10.1017/S0952523808080826

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 679-698.

Cooper, E. A., & Norcia, A. M. (2014). Perceived Depth in Natural Images Reflects Encoding of Low-Level Luminance Statistics. Journal of Neuroscience, 34(35), 11761--11768. http://doi.org/10.1523/JNEUROSCI.1336-14.2014

Devlin, K. (2002). A review of tone reproduction techniques, (November), 1–13.

Drago, F., Myszkowski, K., Annen, T., & Chiba, N. (2003). Adaptive Logarithmic Mapping for Displaying High Contrast Scenes. Computer Graphics Forum, 22(3), 419–426. http://doi.org/10.1111/1467-8659.00689

Field, D. J., & Brady, N. (1997). Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes. Vision Research, 37(23), 3367–3383. http://doi.org/10.1016/S0042-6989(97)00181-8

Ghodrati, M., Morris, A. P., & Price, N. S. C. (2015). The (un)suitability of modern liquid crystal displays (LCDs) for vision research. Frontiers in Psychology, 6(MAR), 1–11. http://doi.org/10.3389/fpsyg.2015.00303

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, *15*(1), 20-25.

Goodrich, G., & Ludt, R. (2002). Change in visual perceptual detection distances for Low Vision travelers as a result of dynamic visual assessment and training. Journal of Visual Impairment & Blindness, 96(1):7-15.

Hassan, S. E., Lovie-Kitchin, J. E., & Woods, R. L. (2002). Vision and Mobility Performance of Subjects with Age-Related Macular Degeneration. Optometry & Vision Science, 79(11), 697–707. http://doi.org/10.1097/00006324-200211000-00007

Kallie, C. S., Legge, G. E., & Yu, D. (2012). Identification and detection of simple 3D objects with severely blurred vision. Investigative Ophthalmology and Visual Science, 53(13), 7997–8005. http://doi.org/10.1167/iovs.12-10013

Krantz, J. H. (2000). Tell me, what did you see? The stimulus on computers. Behavior Research Methods, Instruments, & Computers, 32(2), 221–229. http://doi.org/10.3758/BF03207787

Kwon, M., & Legge, G. E. (2011). Spatial-frequency cutoff requirements for pattern recognition in central and peripheral vision. Vision Research, 51(18), 1995–2007. http://doi.org/10.1016/j.visres.2011.06.020

Larson, G. W., & Shakespeare, R. (2004). Rendering with Radiance: the art and science of lighting visualization. Booksurge Llc.

Legge, Gordon E. (2007) Psychophysics of reading in normal and Low Vision. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Leat, S. J., & Lovie-Kitchin, J. E. (2006). Measuring mobility performance: Experience gained in designing a mobility course. Clinical and Experimental Optometry, 89(4), 215–228. http://doi.org/10.1111/j.1444-0938.2006.00050.x

Loomis, J. M. (2003). Visual space perception: Phenomenology and function. Arquivos Brasileiros de Oftalmologia, 66(5 SUPPL.), 26–29. http://doi.org/10.1590/S0004-27492003000600004

Loomis, J. M., Da Silva, J. a, Fujita, N., & Fukusima, S. S. (1992). Visual space perception and visually directed action. Journal of Experimental Psychology. Human Perception and Performance, 18(4), 906–921. http://doi.org/10.1037/0096-1523.18.4.906

Mather, G., & Smith, D. R. R. (2000). Depth cue integration: stereopsis and image blur. Vision Research, 40, 3501–3506. http://doi.org/10.1016/S0042-6989(00)00178-4

Thompson, W. B., Legge, G. E., Kersten, D. J., Shakespeare, R. A., & Lei, Q. (2017). Simulating visibility under reduced acuity and contrast sensitivity. *JOSA A*, *34*(4), 583-593.

Timmis, M. A., & Pardhan, S. (2012). Patients with central visual field loss adopt a cautious gait strategy during tasks that present a high risk of falling. Investigative Ophthalmology and Visual Science, 53(7), 4120–4129. http://doi.org/10.1167/iovs.12-9897

Matthis, J. S., Barton, S. L., & Fajen, B. R. (2015). The biomechanics of walking shape the use of visual information during locomotion over complex terrain. Journal of Vision, 15(3), 10. http://doi.org/10.1167/15.3.10.doi

Oberholzer, M., Ostreicher, M., Christen, H., & Brühlmann, M. (1996). Methods in quantitative image analysis. Histochemistry and Cell Biology, 105(5), 333–355.

Potter, M. C. (2012). Recognition and memory for briefly presented scenes. *Frontiers in psychology*, *3*, 32.

Rand, K. M., Tarampi, M. R., Creem-Regehr, S. H., & Thompson, W. B. (2011). The importance of a visual horizon for distance judgments under severely degraded vision. Perception, 40(2), 143–154. http://doi.org/10.1068/p6843

Rand, K. M., Tarampi, M. R., Creem-regehr, S. H., & Thompson, W. B. (2012). The influence of Ground Contact and Visible Horizon on Perception of Distance and Size under Sevrely Degraded Vision. Seeing Perceiving, 25(5), 425–447. http://doi.org/10.1163/187847611X620946

Sikl, R., & Simecek, M. (2015). Visual space perception at different levels of depth description. Attention, Perception and Psychophysics, 77(6), 2098–2107. http://doi.org/10.3758/s13414-015-0917-2

Steinicke, F., Bruder, G., & Kuhl, S. (2011). Realistic perspective projections for virtual objects and environments. ACM Transactions on Graphics, 30(5), 1–10. http://doi.org/10.1145/2019627.2019631

Sun, H.-J., Campos, J. L., & Chan, G. S. W. (2004). Multisensory integration in the estimation of relative path length. Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale, 154(2), 246–54. http://doi.org/10.1007/s00221-003-1652-9

Timmis, M. A., & Pardhan, S. (2012). Patients with central visual field loss adopt a cautious gait strategy during tasks that present a high risk of falling. Investigative Ophthalmology and Visual Science, 53(7), 4120–4129. http://doi.org/10.1167/iovs.12-9897

Vilankar, K. P., Golden, J. R., Chandler, D. M., & Field, D. J. (2014). Local edge statistics provide information regarding occlusion and nonocclusion edges in natural scenes. *Journal of vision*, *14*(9), 13-13.

Ward, G., & Shakespeare, R. (1998). Rendering with radiance: The art and science of lighting visualization. San Francisco, CA: Morgan Kaufmann Publishers, Inc