Depth discrimination from occlusions in 3D clutter

Haomin Zheng

Master of Science

School of Computer Science

McGill University

July 2016

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

MCGILL UNIVERSITY

Abstract

School of Computer Science

Master of Science

by Haomin Zheng

Objects such as trees, shrubs, and tall grass consist of thousands of small surfaces that are distributed over a 3D volume. A natural visual task in 3D cluttered scenes is to estimate the depth of objects that are embedded within the clutter. For example, one might estimate the distance to a predator or prey, or decide if a fruit is reachable. To estimate depth in 3D clutter, a visual system can use binocular disparity and motion parallax cues, but these depth cues are less reliable in 3D clutter because surfaces tend to be partly occluded. However, occlusions are not necessarily a nuisance for depth perception in 3D clutter, since occlusions themselves provide depth information. It is unknown whether visual systems can use occlusion cues in 3D clutter, though, as previous studies have considered occlusions for simple scene geometries only. Here we present a set of depth discrimination experiments that examine depth from occlusion cues in 3D clutter. We identify two probabilistic occlusion cues. The first one, *visibility cue*, is based on the fraction of an object that is visible, and the second one, *range cue*, is based on the depth range of the occluders. We show the visual system uses both of these occlusion cues. We also define ideal observers that are based on these cues, and show that human observer performance is close to ideal using the visibility cue but far from ideal using the range cue. The reason is that the range cue itself depends on depth estimation of the occluders from binocular stereopsis or motion parallax cues which is less reliable in 3D clutter. Our results thus provide fundamental constraints on the information that is available from occlusions in 3D clutter, and show how the visual system can discriminate depth by combining these occlusion cues with stereo and motion cues.

MCGILL UNIVERSITY

ABRÉGÉ

School of Computer Science

Master of Science

by Haomin Zheng

Les objets tels que les arbres, arbustes et herbes hautes sont composs de milliers de petites surfaces distribuées dans un volume 3D. Une tâche visuelle naturelle est d'estimer la profondeur des objets dans une scène encombrée de surfaces en désordre. Par exemple, un humain peut estimer la distance d'un prédateur ou dune proie, et peut décider si un fruit, parmi le feuillage, est accessible. Pour estimer la profondeur dans un fouillis 3D, un système visuel peut utiliser la disparité binoculaire et des mouvement parallaxes, mais ces indices de profondeur sont moins fiables dans un fouillis 3D parce que les surfaces sont souvent partiellement occlus. Par contre, les occlusions ne nuisent pas ncessairement à la perception de profondeur dans un fouillis 3D, puisque les occlusions eux-mêmes fournissent de l'information sur la profondeur. Des études précédentes ont seulement examié l'occlusion des géométries dans des scènes simples, mais nous ne savons pas si les systèmes visuels peuvent utiliser les indices fournis par les occlusions pour déterminer les profondeurs dans un fouillis 3D. Ici, nous présentons un ensemble d'expériences qui s'agissent de différencier la profondeur en utilisant les indices visuels fournis par les occlusions dans un fouillis 3D. Nous identifions deux occlusion repères probabilistes. La première est basée sur la partie d'un objet qui est visible, *les indices de visibilité*, et la seconde, *la gamme d'indices*, est basée sur la gamme de profondeur des obstructeurs. Nous démontrons que le système visuel utilise les deux indices d'occlusion. Nous définissons également les observateurs idéaux, basés sur ces indices, et nous montrons que la performance de l'observation humaine est proche de l'idéal utilisant les indices de visibilité, mais elle est loin d'être idéale utilisant la gamme d'indices, la raison étant que la gamme d'indices dépend elle-même de l'estimation de la profondeur des obstructeurs, des indices provenant de la vision stéréoscopique binoculaire ou des mouvements parallaxe qui sont moins fiables dans un fouillis 3D. Ainsi, nos résultats fournissent des contraintes fondamentales sur l'information disponible provenant des occlusions dans un fouillis 3D, et montrent comment le système visuel peut distinguer la profondeur en combinant les indices d'occlusion et les indices stéréos et de mouvements.

Acknowledgements

I would like to thank my thesis supervisor, Prof. Michael Langer. He has been guiding and helping me on every step of my research experiments, and provided invaluable advices for my writing. His courses were also very inspiring and laid down foundation for my study and research in the area of human visual perception. His passion for his research was there to accompany me for every moment and I'm certain it will inspire many more in the future.

A special thanks goes to my former colleague Shayan Rezvankhah. His work was the foundation of all my research towards this thesis, without it I would not be able to achieve this far.

I would also like to thank the School of Computer Science for granting me Graduate Excellence Fellowship. This was not only a great honour, but also provided valuable assistance financially.

Last but not least, I would like to thank my friends, family and loved ones, who supported me through all the expected and unexpected turns of my life. Your love is what made me today, for that I am eternally grateful.

Preface

This thesis revisits and builds upon experiments and theory that were introduced in the M.Sc. thesis ”Depth discrimination in cluttered scenes using fishtank virtual reality” by Shayan Rezvankhah [1]. With similar experiment setup, this thesis focuses on two occlusion cues. By utilizing different distributions for 3D clutter, we are be able to isolate and study occlusion cues in more detail. We also developed ideal observers to run the same tests as human subjects. This also provided new insight into how well occlusion perform in theory.

This thesis used material from ”Depth discrimination from occlusions in 3D clutter”, submitted to Journal of Vision in 2016, of which the thesis author is the second author. Michael S. Langer, the supervisor of the thesis author, is the first author of this paper. In particular, Abstract, Section 1.1, Section 3.2 to 3.5 and Chapter 6 are taken almost directly. Section 1.2, 2.3, 4.1, 4.3.1 and 4.4.3 incorporated some material from above paper. The remaining sections of the thesis are entirely new.

The french version of the abstract is translated by a dear friend, Xoey Zhang.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Previous Work

The human visual system has evolved over millions of years, predominantly in cluttered 3D environments such as forest and grassland. Such environments consist of objects such as trees, shrub, and tall grass that contain thousands of individual surfaces scattered in 3D space. One of the common challenges a human would face in this environment is to estimate depth of a particular target occluded in this clutter scene. When the targets are fully visible, a human uses visual cues such as binocular disparity and motion parallax for depth perception. For the last half century, numerous studies have been conducted, most of them using random dot stereo-grams, to study the underlying mechanics and to measure performance of human depth perception [2] [3] [4]. It also has been shown that humans are the most sensitive to sinusoidal depth modulations with a spatial frequency between 0.2 and 0.5 c/deg, and this is true for either binocular disparity or motion parallax [5]. These experiments suggest that when the targets are fully visible, these two visual cues are the primary source of depth information, and they both utilize the differences in images from different viewpoints. For simplicity, we will refer to both stereopsis and motion parallax cue as *parallax cues* in this thesis.

If we put the target of interest inside a cluttered 3D environment, previous work has shown that such 3D clutter leads to the target being partly occluded, which reduces visibility [6] and complicates the task of depth perception [7]. Because it is very difficult, sometimes even impossible, to find correspondences between images from different viewpoints, the parallax cues would fail. Studies have shown that we are still able to estimate depth, by using the ”Da Vinci” stereopsis cue [8]. By observing an unpaired part of the images, we can estimate depth qualitatively [9]. This shows that our depth perception of targets is mostly hindered by the existence of the clutter, but no study has shown how much this effect is. In this thesis, we will show that even with the occlusions and reduced visibility, our depth perception is still fairly accurate because we are also using the clutter as a source of depth information.

Previous studies of depth perception in 3D clutter have also concentrated on the depth of the clutter elements themselves. For example, many studies using random dot or random line stimuli have asked how many discrete depth planes can be perceived using binocular disparity [10] [11] or motion parallax [12]. Others have asked how well the visual system can perceive the depth to width ratio of the 3D clutter [13] [14]. These studies have provided key insights into depth perception from parallax cues in 3D cluttered scenes. However, we argue that these studies are incomplete since they use only points or lines for the clutter rather than surfaces. As such, these studies neglect occlusion effects which are very important in 3D cluttered scenes.

1.2 Thesis Outline

In this thesis we consider 3D cluttered scenes that consist of 2D surface elements that are randomly placed in a volume and that produce a significant amount of occlusion. We examine the depth cues that are provided by the occlusions and how observers use these cues. Specifically, we examine how well observers can discriminate the depths of identifiable objects that are located within the 3D clutter.

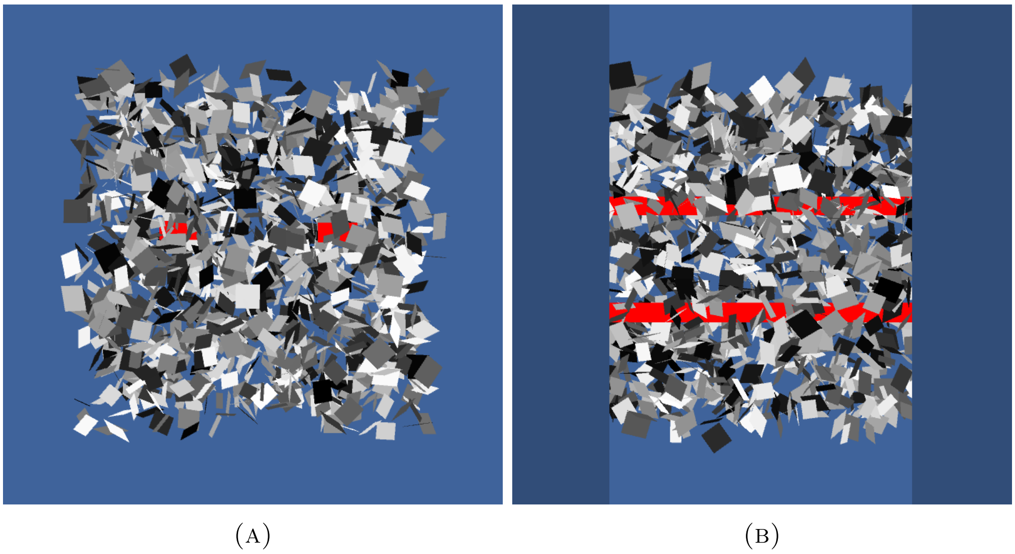


FIGURE 1.1: Each rendered scene consist of two rectangular red targets that are embedded in a 3D field of grey distractors and separated in depth. The targets are (A) short bars separated horizontally or (B) long bars separated vertically. The target surfaces always face the Z direction. The short bar targets have a horizontal:vertical aspect ratio of 2:1. The long bar targets have an aspect ratio of 20:1 and extend beyond the width of the clutter. The left and right edges of the long bar targets are hidden behind large flanking vertical occluders to remove binocular disparity and motion parallax cues.

Consider the examples shown in Figure 1.1. These scenes consist of large number of surface elements (distractors) which define the clutter and two identifiable objects (targets). The clutter consists of random grey colored square distractors that are distributed randomly over a cube volume. The two targets are red rectangles that lie at different depths within the clutter. In Figure 1.1a, the two targets are short bars that are positioned in the left and right halves of the volume. In Figure 1.1b, the targets are long bars that are positioned in the upper and lower halves of the volume.

We present experiments to both human subjects and ideal observers. We study the depth information that is available from occlusions, and how occlusion cues might be combined with binocular stereo and motion parallax cues. We identify two occlusion-based depth cues in 3D clutter. Unlike the classical depth from occlusion cue which defines an ordinal constraint only, the two new occlusion cues are metric depth cues which are based on a probabilistic model of 3D clutter.

Our experiments are conducted using Virtual Reality systems. The idea of VR was first introduced by Sutherland [15] when he developed a research prototype of a head-mounted display. The underlying motivation in virtual reality is to realistically represent 3D virtual worlds to users so that he or she perceives and interacts with them naturally. We first used Oculus Rift DK2 for our experiments, because of its convenience to implement and it has also been shown that this device can provide accurate 3D perception as well as simulate motion parallax [16]. We then repeated our experiments on a Fish Tank VR setup, composed of nVidia 3D display and a head tracking device. It also has been shown that this setup was a drastic improvement upon 3D display without dynamic head coupled perspective [17].

We then proceed to use the same experiment setup for ideal observers. We programmed two algorithms each for one cue, in order to investigate how much information is available to the test subjects. We also made comparisons with different combination of perspective cue. These results will show how the visual system can combine information from occlusion cues with other cues like parallax cues and perspective cues.

Chapter 2

Defining the Occlusion Cues

2.1 Introduction

In a depth discrimination task, it has been shown that human can do extremely well with fully visible targets, but it is expected that if the targets are inside a densely cluttered scene, the performance will become worse. This is because we are relying on traditional cues such as binocular disparity and motion parallax to estimate depth, and occlusion of targets' vertical edges will severely hinder our performance. But in reality, and also in our experiment later, we observe that humans can still do reasonably well with targets in 3D clutter comparing to fully visible targets. We theorize that there are two additional cues that are provided by the clutter, and they are the cues human rely on to do depth discrimination task in a cluttered scene.

To establish these two occlusion cues, we used Unity3D to generate a scene with uniformly distributed occluders and a single red target, as shown in Figure 2.1. We then investigate what kind of information is available by analysing the produced image to determine the correct depth of the target.

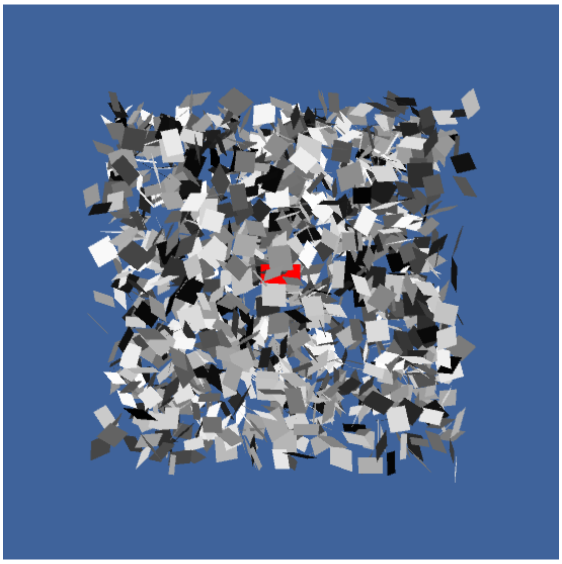


Figure 2.1: Each scene contains 1331 opaque square occluders, each occluder is 1.2cm x 1.2cm in dimension, and its orientation is randomized. All occluders are randomly distributed in a 20cm x 20cm x 20cm axis-aligned cube. We have a single viewpoint pointing towards z axis at the center of the cube, and is 60 cm away from the cube's front face. A target of size 2.4cm x 1.2cm is generated at desired depth in the occluder cube, and it is always facing z axis.

2.2 Visibility Cue

When observing targets in a cluttered scene, the visual system can easily estimate the visible area of the targets. In previous work we can see that if the clutter is uniformly distributed, the visible area of a target decreases as it moves away from the viewing point [7]. This is in part caused by perspective, because the perspective viewing makes closer targets appear larger. But in this particular case, this may also be caused by the clutter distribution, which makes the target appear further away the more it is occluded.

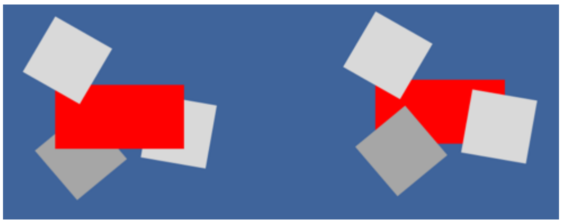


Figure 2.2: As a targets depth increases, the target tends to be more occluded and hence less visible. Target depth refers to the relative depth between the red target and the front face of the clutter cube.

To show that there is a correlation between depth and visible area, we plot the average number of visible pixels of the target across 10000 trials on various simulated depths, both when it is non-occluded and when it is inside the clutter scene shown in Figure 2.1.

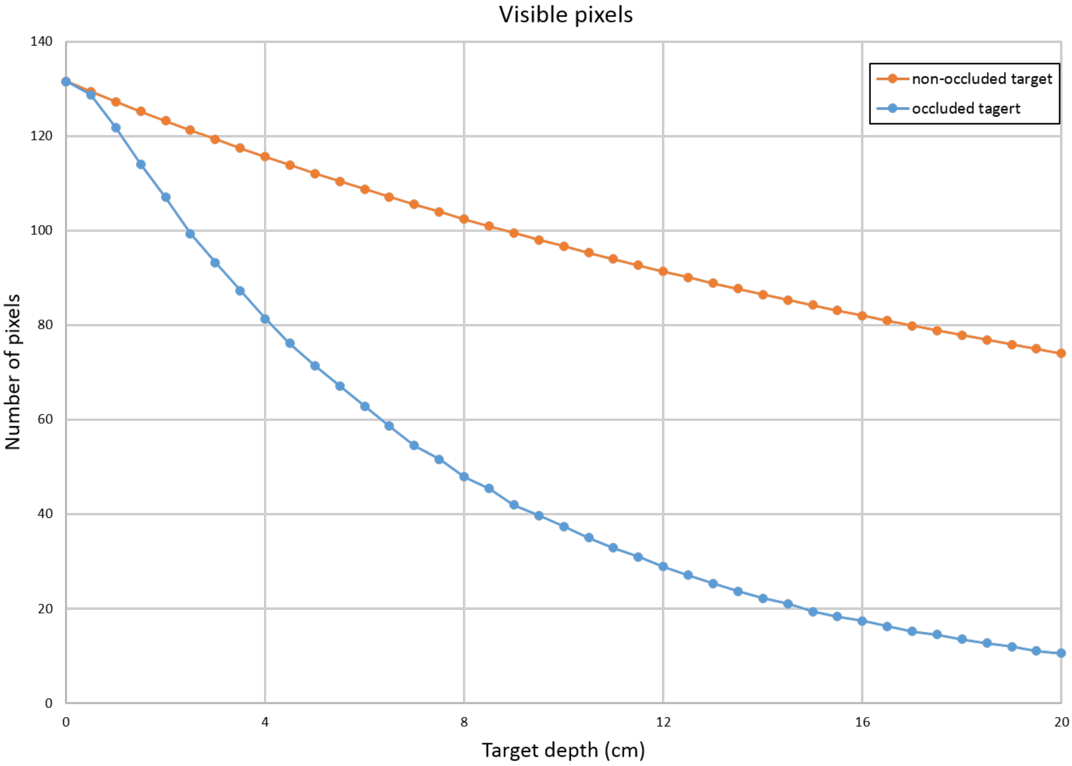


Figure 2.3: Visible area on different distances. Each data point is an average of 10000 trials.

It is already well known that perspective is a reliable cue to depth. Now we would like to investigate whether in the absence of perspective, humans will use the visible area of the target as a reliable cue. To show that this information is indeed consistent, we measure the average visible percentage of pixels for various depths and its standard deviation. The result is shown in Figure 2.4.

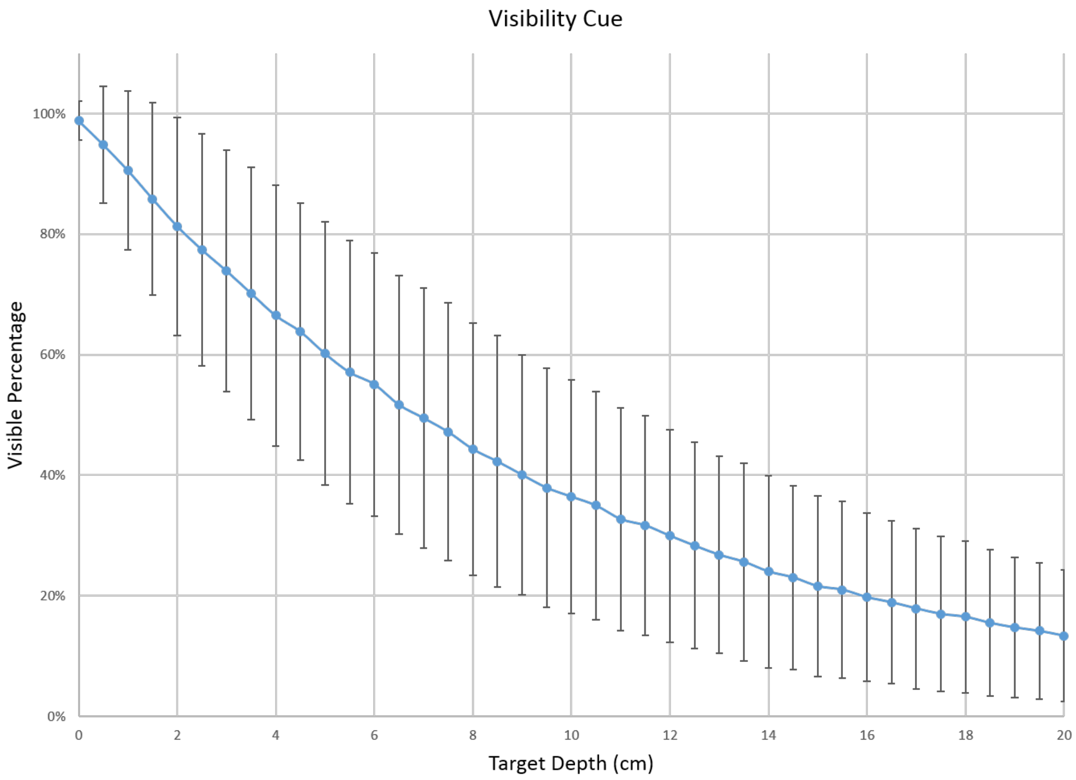


FIGURE 2.4: Visible percentage of pixels of different target depths. Each data point is an average of 10000 trials, error bars indicate standard deviations.

Here we formally define the *visibility cue* as the fraction of area that is visible for depth discrimination. By using the fraction we state that this cue will not use any information provided by perspective.

As we can see, the average fraction of visible area is very reliable but the error caused by randomization of the clutter is very significant. Later in the experiments we will further investigate whether human subjects can use it in practice. Because the perspective cue is already well established and not the focus of this thesis, in all experiments following this one, unless stated otherwise, we will eliminate the size cue by resizing the targets so that they will look identical in size.

2.3 Range Cue

We hypothesize that there is another cue based on a probabilistic relationship between the depth of a target and the range of depths of the surfaces that occlude the target. If the depth of the occluders can be estimated, either using binocular disparity and/or motion parallax cues, then these occluder depths provide a lower bound on the depth of the target. For example, consider two targets that are each partly occluded by one occluder, such as shown in Figure 2.5. Suppose the left target's occluder is at depth 60 cm and right target's occluder is at depth 70 cm. Given only this information, the observer should infer that the right target is more likely to be further, since the left target could lie at depths from 60 to 70 cm (or beyond) but the right target must lie a depth beyond 70 cm. More generally, if an observer can perceive the depth ranges of the occluders of each of two targets, then the observer should infer that the deeper target is the one with the deepest occluder. From now on we will refer to this cue as *range cue.*

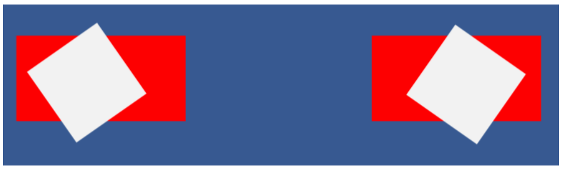


FIGURE 2.5: Illustration of range cue.

To show that this cue is also consistent enough, we measure the depth of the furthest visible occluder that directly occludes the target, and plot the average result for various depths, as shown in Figure 2.6. As we can see, the occluder always has a shallower depth than the target, but on average it is still a very good estimator of the depth of the target. We need to note that this result is based on the assumption that subjects can estimate depth of each occluder accurately. Whether human subjects can use the range cue in practice still remains to be proven in experiments.

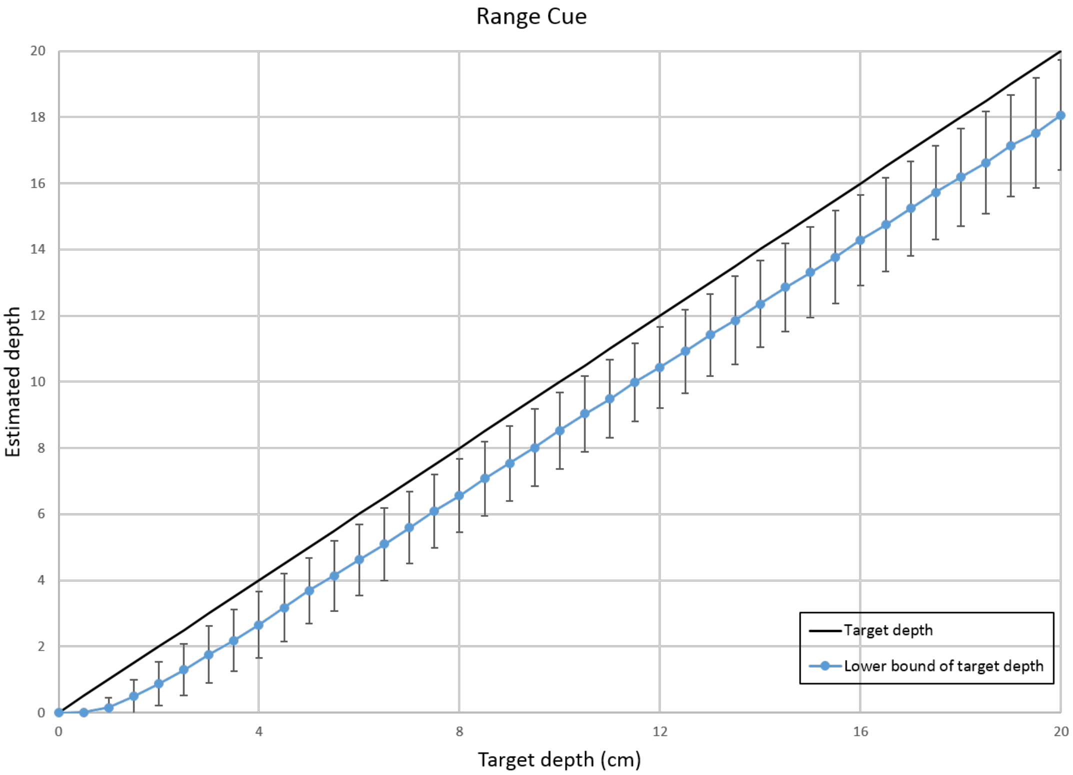


FIGURE 2.6: Range cue by estimating lower bound of target depth from furthest occluder. Each data point is an average of 10000 trials, error bars indicate standard deviations.

Chapter 3

Designing the Experiment

3.1 Motivation and Goal

In a previous study of the same area [1], a psychophysics experiment was designed to measure human subjects' performance for a depth discrimination task in a cluttered 3D scene, but the visibility cue was not tested properly. They tried to remove the visibility cue by removing the clutter near the line of sight to the targets, forming 3D tunnels, but in the end could not provide conclusive evidence as to whether or not the visibility cue was used. In this thesis, we will build upon the same psychophysics experiment setup, and design specific conditions that can isolate the visibility cue. In addition, we will also design conditions with different combinations of the range cue, which was not discussed in any previous studies in this area.

3.2 Apparatus

The rendering and control software ran on a Dell Precision T7610 equipped with an NVIDIA Quadro 4000K graphics card. Scenes were rendered in real-time using VR systems with a head coupled perspective model of the observers left and right 3D eye positions [15]. The systems can track head movement in real time and display the virtual scene rendered from calculated positions of two eyes, in this way the subjects are able to observe the virtual scene from multiply viewpoints and have a consistent perception. Binocular stereo and motion parallax cues were enabled separately based on condition designs, which will be stated in detail in Section 3.3.3.

We used two different display systems. The first system was an Oculus Rift DK2. For this display, scenes were rendered in stereo using 3D Unity, and using C# as the scripting language. The Rift comes with a motion sensing camera and accelerometer for position tracking, and a gyroscope and magnetometer for orientation tracking. For a description of the Oculus Rift DK1 version, see [18]. Observer position tracking was achieved using the Unity plugin provided in Oculus Rift SDK. The position update rate for the Rift is 60 Hz and the display refresh rate is 75 Hz.

The Rift has an OLED display with a resolution of 1920 x 1080 pixels (960 x 1080 per eye), and a nominal horizontal field of view of approximate 100 degrees. This yields about 11 pixels per degree, or 5.5 arcmin separation between pixels. This resolution is relatively low as individual pixels are visible in this display. Moreover, chromatic separation of individual pixels is common. Such limitations would make this display unsuitable for precise depth discrimination experiments but for our task, this resolution was sufficient. To be sure though, we repeated the experiment with a higher resolution display.

The second display was a 1080p Acer GD235HZ stereo monitor (23.6) viewed through NVIDIA 3D Vision shutter glasses. At viewing distance of 60cm, the interpixel distance of the screen was 1.55 arcmin. The screen was refreshed at 120 Hz, so the frame rate for each eye was 60 fps. Scenes were rendered using OpenGL. To render the scene using head coupled perspective, we used the fishtank VR method [17]. We tracked head position and orientation using a mid-range 3D Guidance trackSTAR transmitter (Ascension Tech) with magnetic sensors which were attached to the two handles of the 3D glasses. The position update rate was 80 Hz. Virtual eye position for rendering was set to be along the line segment connecting the two sensors. For the binocular disparity conditions, an interocular distance of 6.5 cm was used. To achieve head coupled perspective, we measured the screen position and orientation relative to the trackSTAR coordinate system which was defined by the magnetic field transmitter. We then combined the emitter, screen, and glasses coordinate systems and rendered each 3D scene in real time from the modelled viewpoint of the viewer.

3.3 Stimuli

A simple illustration of the viewer and stimuli setup is shown in Figure 3.1. Scenes were rendered either with or without binocular disparity cues, and we refer to these as stereo or mono conditions. For the experiment that used the Oculus Rift display, the mono condition presented the same image to both eyes in each frame. The image was rendered from the mid-point between the two eyes. For the experiment that used the fishtank VR display, the mono condition presented an image to one eye only. This was achieved by rendering both eye views, but inserting a large black sphere as a virtual eye patch in front of the scene for one eye, chosen randomly.

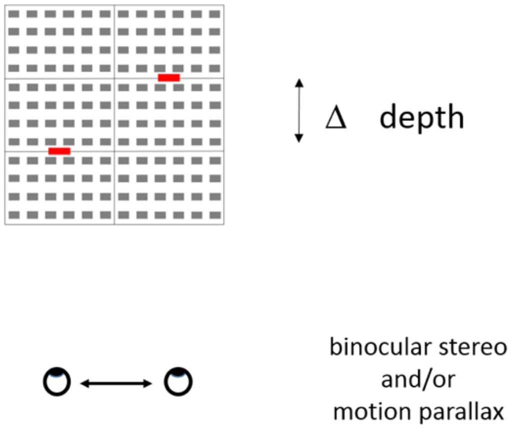


Figure 3.1: The top view of the viewer and stimuli setup. The relative distances and sizes are not to scale. In the experiment, the clutter is a cube volume with sides of length 20 cm and the front of the clutter is at a distance 60 cm for the standard observer (no motion). The uniform grid of distractors illustrates that a uniform probability distribution was used. In the actual scene, occluders are randomized both in position and orientation.

Scenes also were rendered either with or without head coupled perspective. We refer to these as motion or no motion conditions, respectively, although note that the depth information in the motion parallax condition is not merely due to motion, but more generally it is due to head coupled perspective. For the no motion conditions, we instructed observers not to move their heads and in the case they moved their heads by some arbitrarily chosen threshold they were presented a warning on screen and the result of that particular trial was discarded. To ensure that no motion information was available for small head movements below the chosen threshold, we turned off the head coupled perspective and we fixed the virtual observer's position and orientation to a standard view, 60cm from the front and center of the clutter volume.

Scenes were generated as follows. See Figure 1.1. On each trial, the XYZ positions of the two targets were initialized to be at the center of a bounding XYZ volume of size 20 x 20 x 20 cm. We define the standard observer position to be 70 cm from the center of this volume. The targets were separated in depth by an interval Z, namely they were positioned at depths:

$$Z_{near} = 70 - \Delta Z/2 \quad Z_{far} = 70 + \Delta Z/2$$

The value Z was chosen using a staircase procedure that will be described below. The short bar targets then were separated horizontally (X) by 10 cm, and the XY position of each was randomly perturbed by up to 1 cm in X direction and up to 1.5 cm in Y direction. The perturbation is to prevent test subjects from comparing target positions between two consecutive trials and use this information to infer depths.

The 3D clutter in each trial was defined by generating $11^3 = 1331$ distractors. Each distractor was a square of width 1.2cm, and was assigned a random grey level and random 3D orientation. The position of each distractor within the XYZ bounding volume was chosen according to one of the four probability distributions which will be defined below.

The 3D clutter was rendered under perspective projection, with various combinations of stereo and motion cues as described earlier. As stated in Section 2.2, we removed the size cue in this experiment. Specifically, the 3D size of each target was rescaled in real time such that the visual angle of each target was constant, namely it was equal to the visual angle the unscaled target at a depth Z=70. For example, each short bar was roughly 1 x 2 degrees. Removing the size cue from the targets is a standard manipulation in depth discrimination experiments e.g. [19].

3.3.1 Clutter design

To examine occlusion cues, we manipulated the distributions of the clutter. The two cues were visibility and range, as described earlier, and we combined these two cues in four ways as follow.

3.3.1.1 With both visibility and range cue

In this basic scenario, all occluders are uniformly distributed in the cube, illustrated in Figure 3.1. Recall Figure 2.4, in a uniformly distributed clutter, the average fraction of the visible area depends on the target depth. Also recall Figure 2.6, the lower bound of target depth is infered from the depth of the furthest occluder, which also depends on target depth. We expect test subjects to be able to utilize both occlusions cues in this condition. Figure 3.2 shows an example of such scene using a stereogram.

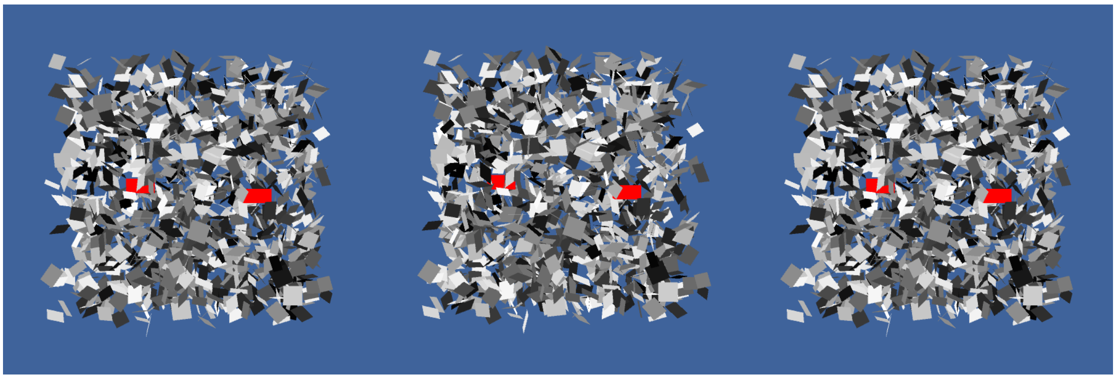


Figure 3.2: Stereogram for the scene with both visibility and range cue, the images are for the right-left-right eye, i.e. the pairs on the left should be cross-fused and the pairs on the right should be viewed divergently. The closer target is on the left, and $\Delta Z = 8$.

3.3.1.2 With neither visibility nor range cue

To eliminate both cues, we modify the distribution so that between the depths of two targets, occluders are not generated, but we keep the total number of occluders the same by making the front/back clutter more dense. This is illustrated in Figure 3.3.

We will now analyse the cues present in this distribution. Because the amount of occluders lying in front of either target are the same, so on average, the visible area of either target should be the same. This means the subjects can not infer any depth information from the visibility cue. Also because the occluders lying in front of targets are of the

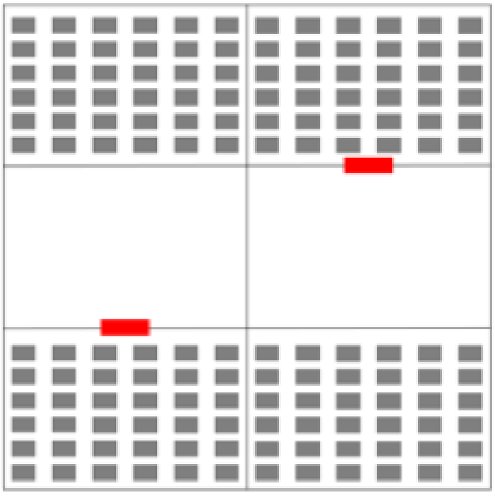


FIGURE 3.3: The top view illustration of the clutter design with neither visibility nor range cue.

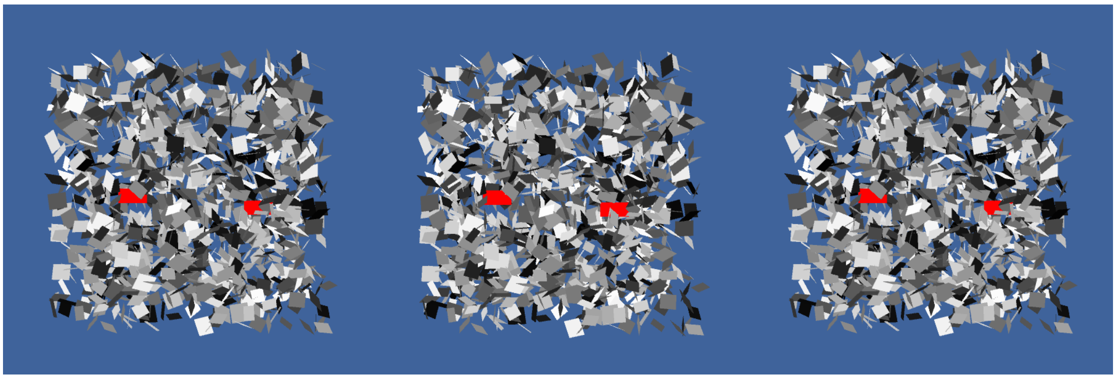


FIGURE 3.4: Stereogram for the scene with neither visibility nor range cue. The closer target is on the left, and $\Delta Z = 8$.

same range of depth, subjects cannot use their depth information to infer the relative target depth either, which eliminates range cue. We expect subjects to perform much worse in this case than with the uniform distribution.

To achieve the same result, we also have a choice of using the same density of the occluders as our uniform distribution, or use the same total number of occluders. We decided against the former because it will make both targets less occluded, leading to stronger parallax cues, making it hard to compare the results.

3.3.1.3 With range cue but without visibility cue

To eliminate visibility cue, the simplest way is to modify the distribution so that the number of occluders in front of a target is always the same no matter the depth of the target. This way, it always has the same expected average visibility as when it is in the middle of the clutter. To implement this, we keep the total number of occluders constant and divide the cube into four sections, left and right, in front of targets and behind targets. Each section has 25% of total occluders. This is illustrated in Figure 3.5.

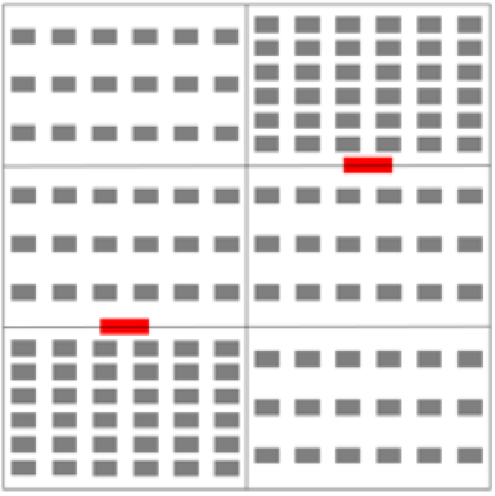


FIGURE 3.5: The top view illustration of the clutter design with the range cue, but without visibility cue.

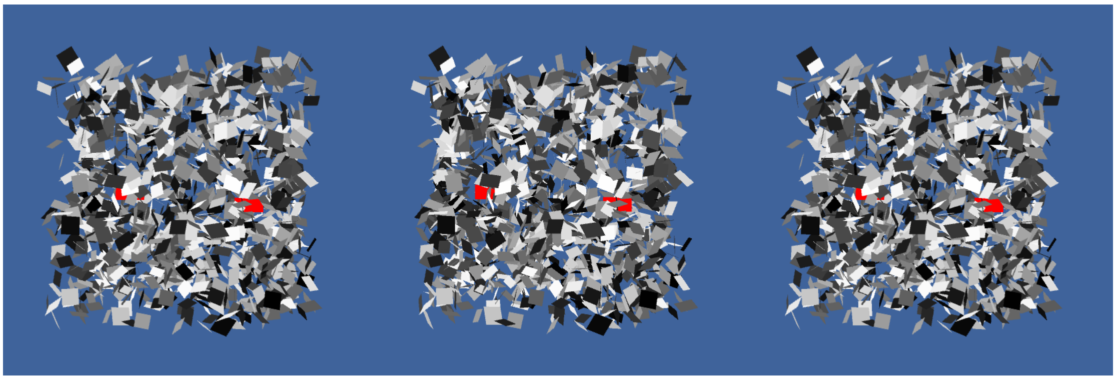


FIGURE 3.6: Stereogram for the scene with range cue, but without visibility cue. The closer target is on the left, and $\Delta Z = 8$.

It is clear that the range cue is still present in this distribution, because the occluders of each target have a different range of depth. Therefore the furthest occluder among each has a different expected depth. We expect subjects to perform worse than uniform distribution because of the absence of visibility cue, but better than the distribution with neither cue.

3.3.1.4 With visibility cue but without range cue

It is quite tricky to eliminate the range cue. Instead of achieving this directly, we started with the second distribution mentioned above, which contains neither cue, and added the visibility cue back in. Taking the said distribution, divide it into four sections, and make the occluders in front of the closer target less dense, and that of the further targets more dense. The densities depends on the depth of both targets so that the number of occluders in front of the targets remains the same as in the case of uniform distribution. This is illustrated in Figure 3.7.

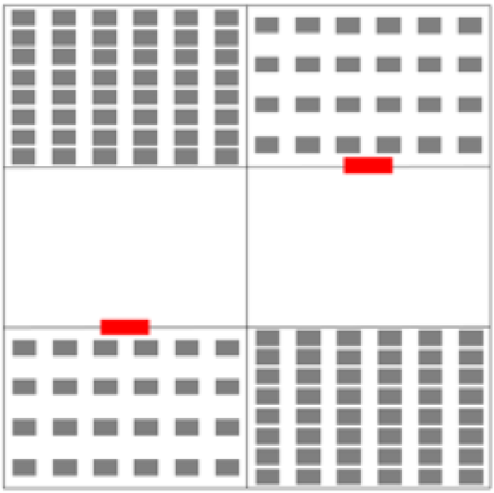


FIGURE 3.7: The top view illustration of the clutter design with visibility cue, but without range cue.

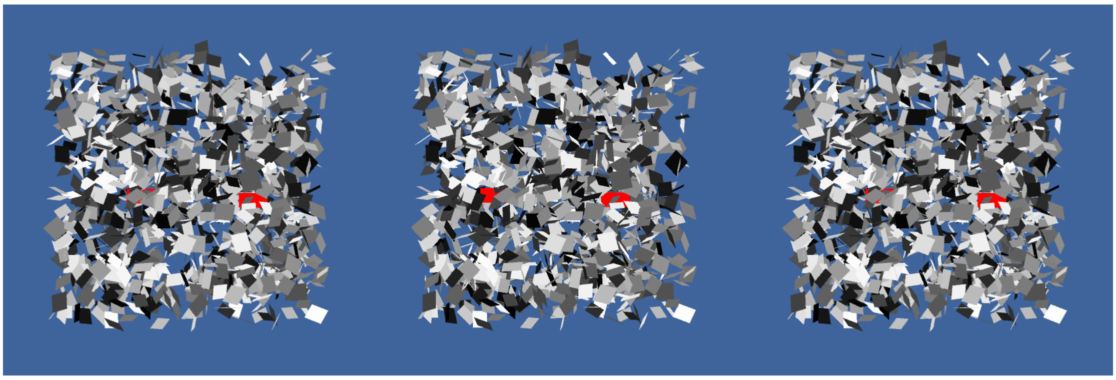


FIGURE 3.8: Stereogram for the scene with visibility cue, but without range cue. The closer target is on the left, and $\Delta Z = 8$.

Because the number of occluders in front of the targets is the same as in the uniform distribution, the visibility cue is unaffected. But because there is still a gap in the depth range between two targets, the front occluders of each target are randomized within the same range, so there is no range cue present. We expect subjects to perform worse than with the uniform distribution but better than the distribution with neither cue.

3.3.2 Target design

In a previous study [1], red squares were used to represent targets, but from preliminary tests we discovered that small red squares are too often completely occluded. We also tried the clutter with less density, but we discovered that both visibility cue and occlusion cue became weaker and less obvious. Because subjects use the targets' vertical edges for binocular and motion parallax, we also do not want to increase their height as this will increase these cues, and perhaps make these cues dominate over the occlusion cues, which are the cues of interest. In the end, we decided to use a rectangle with a width height ratio of 2:1.

Even though the design of small targets has more similarity with objects in natural scenes, e.g. a fruit in a tree full of leaves, a prey hiding in the bush, we are also interested to see how subjects will perform without parallax cues like binocular and motion parallax. To do this, we need to make the vertical edges not visible to the subjects. We then modified our targets into long bars, both ends extending outside of the clutter cube. To make sure the subjects are not seeing the targets' vertical edges, we put two larger rectangular occluders in front of the clutter cube, covering both ends of the targets outside of the clutter. The long bar targets can no longer be placed side by side, so we made them above and below, 6.67 cm (1/3 of the clutter height) away from each other and from the top down edges. Figure 3.9 shows long bar targets with all four clutter design stated above.

To discriminate the depth of these long targets, the subjects have to use the visibility cue and occlusion cue. Also because visual area of the targets increased about seven times, we expect the effect of both cues to be more consistent. But it is unclear whether the performances for long targets will exceed that of short targets. We decided to keep both short and long targets because the results in both cases can provide substantial insight into occlusion cues.

3.3.3 Combination of conditions

Four depth cues (binocular disparity, i.e. stereo, motion parallax, and the two occlusion cues) and two different targets were combined to give 32 possible conditions. For several of these combinations, the task was impossible. We did not test these conditions. For the short bar targets, there were two impossible conditions, namely when there was no stereo, motion, and visibility cues (with or without a range cue). The task is impossible for the condition with the range cue alone (Figure 3.7 with no stereo or motion) because that range cue requires that the observer can perceive the depth of the occluders to some extent, which requires either stereo or motion. For the long bar targets, the task was impossible for the two conditions just mentioned. In addition, the task was impossible when both the visibility and range cues were removed, even if there were stereo or motion cues. The reason is that the stereo and motion cues provide depth information only about the distractors, but not about the targets.

In addition to the 25 conditions stated above, we also tested a 'baseline' condition for the short bar targets in which stereo and motion cues were present, but the cluttered distractors were removed. This gave a total of 26 conditions. We did not include a baseline condition for the long bar targets since this condition was designed to not have any stereo or motion cues about target depth, so it would be impossible for the subjects to estimate the depth.

Each observer ran all 26 conditions in a blocked design, with one staircase per block. The staircases will be described below. The ordering of the blocks was randomized for each subject.

3.4 Test subjects

15 subjects participated in the experiment using the Oculus Rift VR system, and a new set of 15 for the fish tank VR system. Each subject was a student at McGill University and was paid $10. Subjects had little or no experience with psychophysics experiments. Each had normal or corrected-to-normal vision. We required that each subject could discriminate 50 arcsec of disparity to participate, namely level 6 of the Randot Stereo Test (Precision Vision). Subjects were unaware of the purpose of the experiments. Informed consent was obtained using the guidelines of the McGill Research Ethics Board which is consistent with the Declaration of Helsinki.

3.5 Procedure

In each trial, subjects indicated which of the two targets was closer to them. They responded by the pressing keys on the keyboard: left-right arrows for short bar targets, and up-down arrows for the long bar targets.

As mentioned above, a blocked design was used such that the combination of cues was fixed for each block. A one-up/one-down staircase was used for each block with different step sizes for down steps versus up steps. The ratio between the log of the up-step size and the log of the down-step size was chosen as 0.2845 [20]. Specifically, whenever the subject answered correctly, we reduced the distance ΔZ between targets by a factor 0.8, and when the observer answered incorrectly we increased ΔZ by a factor 2.19. This ratio aims for approximately 78 percent correct. Each staircase began at $\Delta Z = 12$ cm and terminated after 12 reversals. To compute the threshold for a given staircase, we averaged the log of the ΔZ values for the last 10 reversals. If ΔZ increased beyond 20 cm which normally would put the targets outside the bounding box of the clutter, we instead displayed the near target just in front of the front face at Zmin and the far target just beyond the back face at Zmax. This configuration made the task trivial since the near target was unoccluded and the far target was highly occluded. If the observer still answered incorrectly in this case, we used the usual rule for choosing the next staircase level but in the next trial we again displayed the targets at the same depths just below Zmin and just beyond Zmax. This gave the same target images anyhow because of how we scaled the target sizes.

For blocks in which there was a motion cue present, subjects were instructed to move their heads left and right. If they did not move, then a warning message was displayed and the trial was discarded. We clipped the rendered observers position to a horizontal

XYZ line segment of size 30 x 0 x 10 cm which was centered at position (0, 0, 60) relative to the center of the front face of the clutter cube. This restricted the viewing position to always have the same Y value, which removed any possibility that the observers could use vertical motion parallax from the targets upper and lower (horizontal) edges. For blocks in which there was no motion cue present, a message was presented telling subjects not to move their heads (see Stimuli section) and the trial at that level was repeated with a new stimulus.

The response time in each trial was limited to four seconds. If the subject didn't respond, then the trial was discarded and another scene was generated using the same target distance. A prompt was displayed to remind the subject to respond in time. The experiment typically lasted close to one hour.

Before running the experiment, each observer ran a short practice session with three conditions, each with stereo present: the short bar targets with and without motion parallax, and the long bar targets with motion parallax. There was no time limit in each trial of the practice session. As in the real experiment, the initial ΔZ was 12 cm and a staircase was used to determine the next level. Since the purpose of the practice session was merely to familiarize the subjects with the requirements of the task, we kept the session short: each condition terminated with the first incorrect answer.

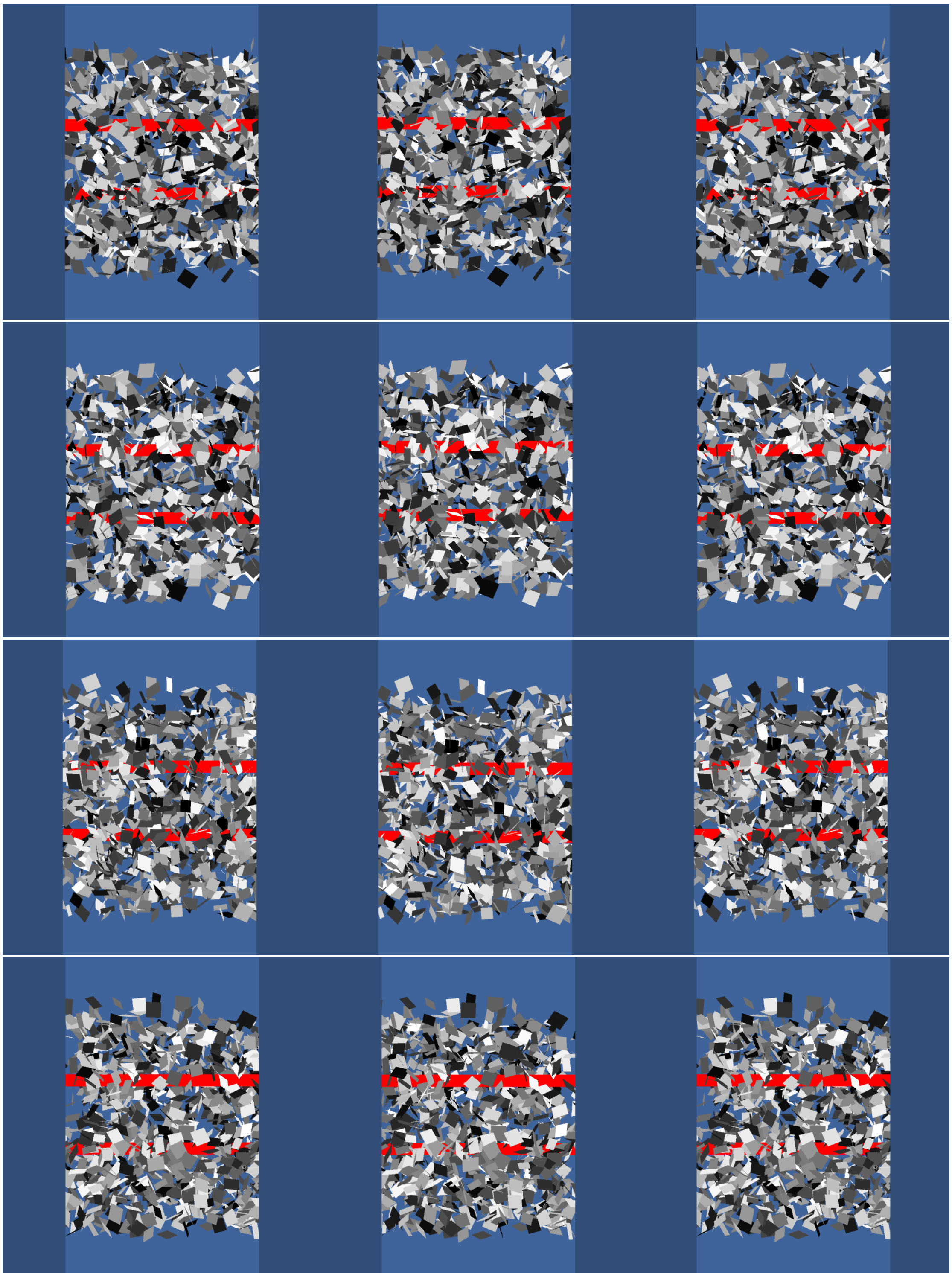


FIGURE 3.9: Stereogram for the scene with long targets. From top to bottom, the distributions are, with both visibility and range cues, with neither cues, with range cue only, with visibility cue only. The closer target is always on the top, and $\Delta Z = 8$.

Chapter 4

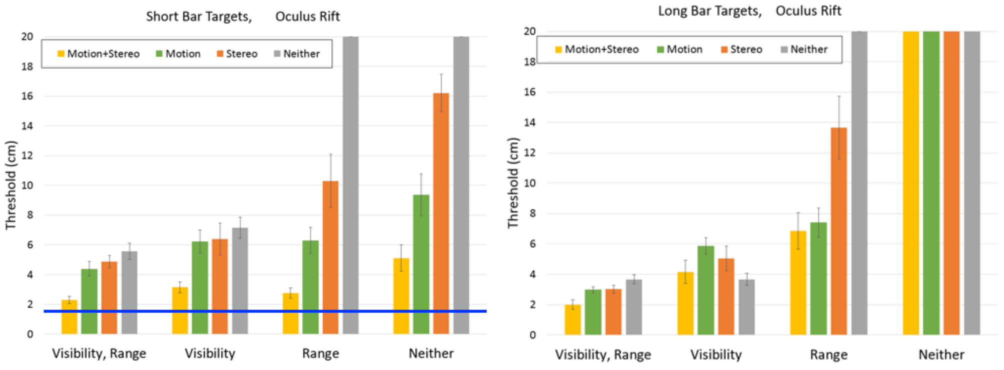
Result and Analysis

4.1 Motivation and Goal

The main goal in these experiments was to examine whether the two types of occlusion cues were used to discriminate depth in 3D clutter and how these cues would interact with binocular stereo and motion parallax cues. If both types of occlusion cue were used by human observers, then we would expect performance to be best when both types of cues are present, and we would expect performance to be better when one of these cues is present than when neither is present. We also expect that, within any of the four combinations of the occlusion cues, performance should be best when both stereo and motion cues are present since previous studies have shown that stereo and motion cues combine to give better performance. Such studies traditionally use scenes containing isolated surfaces [21] [22] and some studies also have used scenes containing 3D clutter [23] [17]. Note that these previous studies have examined interactions between stereo and motion cues and occlusion cues explicitly, which was the goal of our experiments.

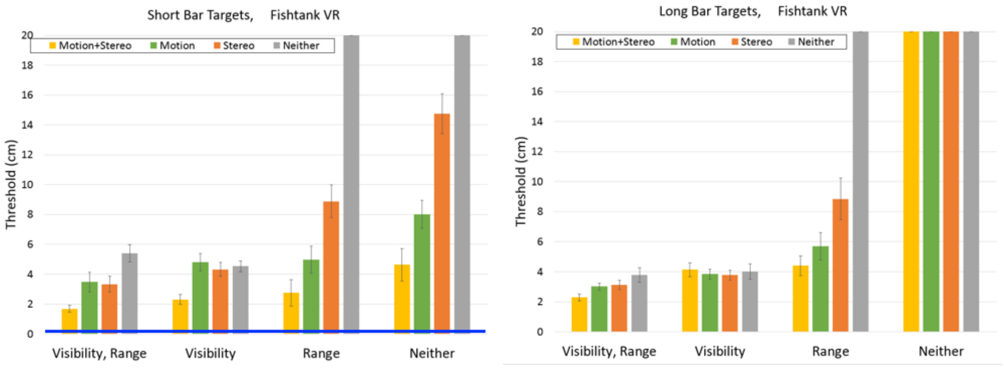
4.2 Results

Figure 4.1 shows the mean of depth differences thresholds (ΔZ) for all 3D clutter conditions using the Oculus Rift setup. Figure 4.2 shows the result of the same experiment but conducted using fish tank VR. For conditions that we did not test, we plotted a threshold of $\Delta Z = 20$ cm. These were conditions in which the task was impossible for the subjects for ΔZ values less than 20 cm and the task was trivial when ΔZ was greater than 20 cm since the near target was out of the occluder distribution cube and fully visible. More discussion on this topic is continued in Section 4.4.1.



(A) (B)

FIGURE 4.1: Result data for experiments done using Oculus Rift. Blue line indicates baseline performance. Lower threshold indicates better performance. A threshold of 20cm was given to conditions we believe impossible for human subjects. Error bars show standard error of the mean.



(A) (B)

FIGURE 4.2: Result data for experiments done using fishtank VR (NVidia 3D display + Trakstar). Blue line indicates baseline performance. Lower threshold indicates better performance. A threshold of 20cm was given to conditions we believe impossible for human subjects. Error bars show standard error of the mean.

Thresholds are given as ΔZ values in cm. To convert these thresholds to stereo disparities, we use:

$$\text{disparity in arcmin} \approx 4.5 \text{ arcmin per cm} * \Delta Z \text{ in cm}$$

For example, the min and max thresholds which are roughly $\Delta Z = 2$ and 20 correspond to 9 and 90 arcmin of disparity respectively. The conversion assumes the observer is at 60 cm from the front face of the clutter.

4.3 Analysing the Importance of Occlusion cues

From the above results, it is easy to see the trend that performance gets worse when we remove either of the occlusion cues and worst when we remove both. We used the following two statistic methods to reinforce this claim.

4.3.1 t-test

To analyze the human observer data in Figure 4.1 & 4.2, we compare thresholds across several conditions. We use paired two-sided t-tests (Microsoft Excel) to test the null hypothesis that the means of two conditions across 15 observers are the same.

We first discuss the results for baseline conditions (shown with blue line). Recall from the last chapter, that in the baseline condition we have only short bar targets with stereo and motion cues but no clutter. For the Oculus Rift display, the mean baseline threshold was $\Delta Z = 1.6$ cm. This corresponds to a binocular disparity of about 7.3 arcmin, which is slightly greater than the nominal interpixel distance of the Oculus Rift display (5.5 arcmin), presumably because the resolution for this display is worse than the nominal value because of chromatic aberrations. For the fishtank VR display, the mean Z threshold for the baseline condition was 0.2 cm which corresponds to a binocular disparity of 0.9 arcmin. This threshold is lower than the interpixel distance for the Acer monitor (1.55 arcmin at a screen viewing distance of 60 cm).

The baseline thresholds were lower than the 3D clutter thresholds for the short bar targets and for the clutter condition in which all four cues were present (Fig. 4.1a leftmost yellow bar). For the Oculus Rift display, the difference was close to significant (t = 1.94, p = 0.07). For the fishtank VR display the difference was highly significant (t = 7.4, p = 0.000003). Thus the presence of clutter indeed reduced performance relative to the baseline.

We next show that, although the 3D clutter reduced performance, observers were still able to use the information in the 3D clutter through the two occlusion cues we hypothesised. We first consider the visibility cue. We compare the ”Visibility” conditions with ”Neither” conditions, and ”Visibility, Range” conditions with ”Range” conditions, these comparisons are done with four different parallax cue combinations, along with two different targets, and two display setups. The t and p value for each test is shown in Table 4.1. Note that because we conducted the experiment only using a selected set of conditions (excluding ones we considered impossible for the subjects), several comparisons will not be shown.

As we can see, most of the differences in Table 4.1 are considered statistically significant. This shows that among most conditions, the subjects will perform better when the scene contains the visibility cue, compared to the corresponding scene where the cue is not available. This serves a significant evidence that the human visual system is indeed using the visibility cue on many occasions.

Visibility Cue Test		Oculus Rift		Fish Tank	
		Visibility vs. Neither	Both vs. Range	Visibility vs. Neither	Both vs. Range
Short Target	Motion + Stereo	t=2.38 p=0.032	t=1.39 p=0.186	t=2.31 p=0.037	t=1.22 p=0.243
	Motion only	t=1.89 p=0.079	t=2.72 p=0.017	t=3.05 p=0.009	t=1.71 p=0.109
	Stereo only	t=6.13 p=2.62E-05	t=3.58 p=0.003	t=7.88 p=1.64E-06	t=5.29 p=0.0001
	Neither	N/A	N/A	N/A	N/A
Long Target	Motion + Stereo	N/A	t=4.89 p=0.0002	N/A	t=3.07 p=0.008
	Motion only	N/A	t=4.65 p=0.0004	N/A	t=3.21 p=0.006
	Stereo only	N/A	t=5.23 p=0.0001	N/A	t=4.51 p=0.0005
	Neither	N/A	N/A	N/A	N/A

Table 4.1: Statistics for difference between performance with visibility cue and without it, using paired two-sided t-tests. Because our experiment used a selected set of conditions, some tests can not be done. Yellow colored cell indicates the difference is statistically significant at p ¡ 0.05 level.

There are four cases where the difference is not considered significant. Other than the fact that the number of our test subjects is limited, we also argue that this may be because the thresholds for these cases are very low, meaning they are performing very well using other cues already. Because the visibility cue is probabilistic in nature, it is much less reliable with lower ΔZ. We will verify this hypothesis in the ideal observer chapter later.

Now we examine the range cue. Similar as above, we compare the ”Range” conditions with ”Neither” conditions, and ”Visibility, Range” with ”Visibility” conditions. These comparisons are also done with four different parallax cue combinations, along with two different targets, and two display setups. For the same reason, several comparisons are not shown here. The t and p value for each test is shown in Table 4.2.

Again, most of the test also indicate a significant difference. We can also come to the conclusion that the human visual system uses the range cue on many occasions. It is important to note that there are several tests that we did not expect a difference in theory (shown with *). These conditions, although they use a clutter distribution that

Range Cue Test		Oculus Rift		Fish Tank	
		Range vs. Neither	Both vs. Visibility	Range vs. Neither	Both vs. Visibility
Short Target	Motion + Stereo	t=2.90 p=0.012	t=2.64 p=0.019	t=3.06 p=0.008	t=2.17 p=0.048
	Motion only	t=2.54 p=0.023	t=2.33 p=0.035	t=2.62 p=0.020	t=1.76 p=0.101
	Stereo only	t=3.78 p=0.002	t=1.56 p=0.142	t=5.57 p=6.95E-05	t=1.44 p=0.171
	Neither	N/A	t=1.92 p=0.076 *	N/A	t=-1.14 p=0.274 *
Long Target	Motion + Stereo	N/A	t=2.47 p=0.027	N/A	t=3.57 p=0.003
	Motion only	N/A	t=4.64 p=0.0004	N/A	t=2.35 p=0.034
	Stereo only	N/A	t=2.43 p=0.029	N/A	t=1.61 p=0.130
	Neither	N/A	t=-0.004 p=0.997 *	N/A	t=0.37 p=0.716 *

Table 4.2: Statistics for difference between performances with range cue and without it, using paired two-sided t-tests. Because our experiment used a selected set of conditions, some tests can not be done. Yellow colored cell indicates the difference is statistically significant. * indicates a case where we do not anticipate a difference, because we expect test subjects not able to make use of range cue even though it is present.

contains the range cue, human observers are not able to make use of it. The reason is because the range cue mechanism requires the observer to estimate the depth of the occluders of the target, but with both stereo and motion parallax absent, i.e. "Neither" row in the table, the human observer can not estimate the occluders' depths at all. As we can see in entries with * in Table 4.2, these tests indeed do not show a significant difference.

4.3.2 Multiple linear regression

In our experiment, we assume there are four different cues that are used, two occlusion cues and two parallax cues, with different combination in each staircase test. Cue combination can be very complicated to model, and here we attempt to model them using a simplified linear model. We assume a binary term (X_{cue}) to indicate whether a cue is present (1) or not (0), and result threshold ($Y_{threshold}$) from each staircase test of each test subject as output.

$$Y_{threshold} = \beta_{intercept} + \beta_{stereo} X_{stereo} + \beta_{motion} X_{motion} + \beta_{visibility} X_{visibility} + \beta_{range} X_{range}$$

We excluded the baseline result because there is no clutter in the scene and parallax cue is extremely powerful in this case, unlike other conditions. We also excluded the results from long target experiments, because there is no parallax cue involved even when stereo or motion parallax is enabled. We used multiple linear regression (Microsoft Excel) to solve for the coefficient (β_{cue}) of each term, the result is shown in Table 4.3.

Cue	Coefficients	t Stat	P-value
Intercept	12.57	24.40	3.84E-82
Stereo	-1.40	-3.68	0.0003
Motion Parallax	-4.09	-10.74	6.55E-24
Visibility	-4.38	-11.50	9.35E-27
Range	-2.14	-5.88	8.57E-09

Table 4.3: Multiple linear regression on all data, using different combinations of cues as input. Coefficients are negative because each cue will increase the performance therefore bring down the threshold, which is input Y for the regression analysis.

As we can see, all four cues are shown to have a very significant coefficient. This further proves that both occlusion cues, visibility and range cue, are used in our experiment and improved the depth discrimination performance in a cluttered scene by a significant amount.

It is important to note that, however, this linear model is naive in nature and cannot represent the complex mechanism of visual cue combination. For instance, range cue is dependent on the fact that observer can estimate depth of the occluders, i.e. there should be a coefficient for (range cue AND stereo cue) OR (range cue AND motion cue), which can not be captured in this linear model. Therefore this model is not sufficient enough to provide quantitative estimate of cue importance, but we can still make qualitative conclusion that both occlusion cues are used by subjects in this experiment and their effects are significant.

4.4 Further Discussion

4.4.1 Results for conditions not included in the experiment

Because of our design of clutter distributions, we expected several combinations of conditions are impossible for subjects to perform when ΔZ <20 cm. (For ΔZ >= 20 cm the task is trivial because the front target is outside of the clutter thus entirely visible.) For these conditions we put 20 cm as placeholders in Figure 4.1 & 4.2. For short targets, this includes distributions without the visibility cue and when neither of the parallax cues (stereopsis and motion parallax) are available. Without these three cues, the only information available is the range cue, but for the range cue to function, observers need to estimate the depths of the occluders, which is impossible without either parallax cue. For long targets, in addition to the combination above, the task is impossible if both occlusion cues are absent. Because long targets do not have visible vertical edges, therefore subjects cannot use either of the parallax cues, so without occlusion cues, they can not estimate the depth at all.

During the preliminary tests, we find that these conditions are frustrating for test subjects and the result is often beyond 20 cm and do not provide interesting data. We decided it was not meaningful enough to run these conditions, this also shorten the overall runtime for the subjects. Instead, we programmed a ”dummy” observer to generate random result for ΔZ <20 cm and correct result for ΔZ >= 20 cm, and uses the same parameters for the staircases. The mean of 15 staircases from this dummy observer is 21.33 cm and standard deviation is 1.96 cm. To simplify the matter, we proceed to use 20 cm as placeholders.

It is important to note that we did not include 20 cm or the result from dummy observer for the purpose of significance tests in the previous section. We felt that these are not actual experiment data and it is unfair to compare generated data with collected data.

4.4.2 Comparing two display setups

If we compare the results from Oculus Rift VR (Figure 4.1) and fish tank VR (Figure 4.2), the latter generally has a better performance across all conditions. Particularly for baseline condition, its threshold is almost 1/10 of Oculus Rift. To analyse this statistically, we used the same multiple linear model described above, and added another input X with 1 indicating Fish Tank and 0 for Rift, this term produced a coeffcient of -1.16 cm with t=-3.22 and P=0.001. This shows that for our experiment, subjects using Oculus Rift have a significantly worse performance.

As we stated before, the vast difference for baseline condition can be explained by interpixel distance in visual angle (5.5 arcmin vs. 1.55 arcmin). This difference is less significant, however, when we added in the clutter distribution and the subjects are using occlusion cues addition to parallax cues. This is because the occlusion cues are using probabilistic model and do not rely on the resolution of the image as much as parallax cues do.

4.4.3 Comparing the two parallax cues

In most cases, the performance between motion parallax and stereo is very similar. This is because the information in the image that is provided by these cues is very similar. However, in the range only conditions, motion parallax performs significantly better than stereo. The reason for this result may be that stereo is less reliable than motion when depth differences are large, because of fusion limits [19] [24]. That is, binocular stereo suffers beyond Panums fusional area, but motion parallax does not.

4.4.4 Comparing results between long targets and short targets

The initial intent of the long targets is to eliminate parallax cues, but this will also increase the effectiveness of occlusion cues. Recall that for short targets both visibility cue and range cue are indeed reliable, but they also have a very high standard deviation. We argue that by expanding the target area for seven times (long target vs. short target), the standard deviation should decrease drastically. To illustrate this, we revisit Figure 2.4 & 2.6, and calculated using long targets as well. The result is shown in Figure 4.3 & 4.4.

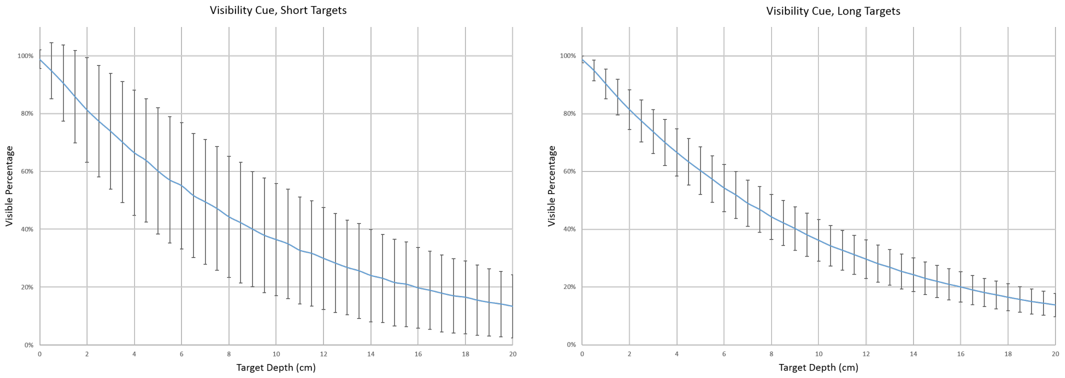


Figure 4.3: Comparing reliability of visibility cue for short targets and long targets. Each data point is an average of 10000 trials, error bars indicate standard deviations.

As we can see, the standard deviation for both cue estimation is lower for long targets than for short targets. Note that this result is based on the assumption that observers can accurately estimate depth of occluders, and this is obviously not true for human observers, but we believe this increase of reliability is applicable for human. This is the reason that, as shown in Figure 4.1 & 4.2, although long bar targets lose both parallax

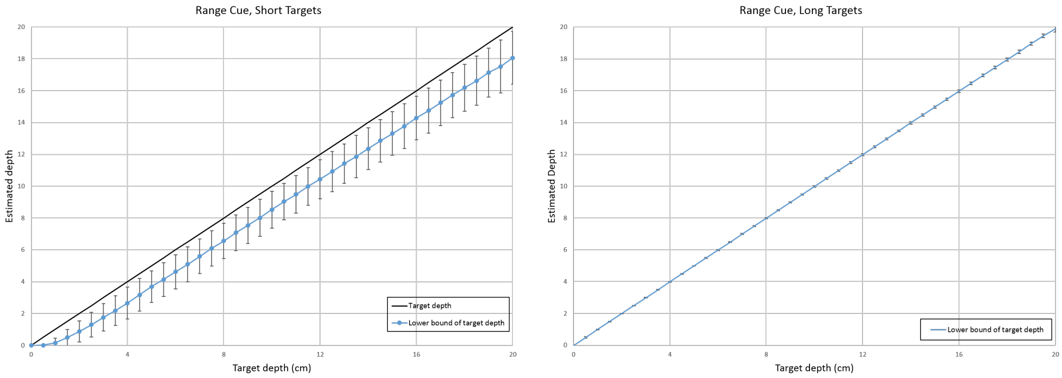


FIGURE 4.4: Comparing reliability of range cue for short targets and long targets. Each data point is an average of 10000 trials, error bars indicate standard deviations. Note that for long targets, error bars are extremely small and hardly visible.

cues, their overall performance is not significantly reduced, because the occlusion cues became more reliable.

4.4.5 Comparing stereo and motion parallax conditions for long targets

Because there is no direct parallax cue for long targets, the parallax conditions (Stereo and/or motion parallax) cannot influence the performance directly. However, we hypothesize that they can still influence the performance of long targets through occlusion cues in the following two ways:

- Parallax cues should improve performance of range cue.

  Recall that the mechanism of the range cue requires the observers to correctly estimate the depth of the occluders of the target. This means that when the range cue is present, performance should be better when either stereo or motion parallax is enabled and best when both are enabled. This is keeping with our result in Figure 4.1b & 4.2b, although not all of the improvements are statistically significant.

- Parallax should improve performance of visibility cue.

  When stereo and/or motion parallax is enabled, the subject can view the target and estimate the visibility from more than one viewpoint. This should reduce the randomness of the visibility cue and make it more reliable compared to the mono and no-motion-parallax condition. However, this is not reflected in our results in Figure 4.1b & 4.2b. For visibility only condition, there is no significant improvement with adding either or both parallax cues. This is further explained in the next chapter with the result from ideal observer, which can show that multiple viewpoints from the observer only provide a marginal amount of information and not enough to improve the performance by a significant amount.

Chapter 5

Ideal Observers

5.1 Motivation

With the results from the human experiments, we can conclude that both the visibility cue and range cue are used in depth perception in 3D cluttered scenes. This means these two cues both have reliable information of target depth. To study how much information these two cues carry, and how much do human subjects make use of them, we developed ideal observer algorithm for each cue. From this we intend to the investigate the limit of information available through either of the occlusion cues.

5.2 Implementation

Because of its simplicity for implementation, we chose to develop ideal observers in Unity3D. We developed two algorithms, one for each cue, that use the information from the scene, such as ray cast hit and depth of objects, to make judgement on the relative depth of the two targets. The detailed algorithms are described below.

5.2.1 Visibility cue algorithm

Our model of a human using the visibility cue is by comparing the visible percentage of the two targets, then choose the larger one as the closer one.

To measure the visibility by algorithm, we first divide the target into a grid of small squares, each the size of 0.5 mm by 0.5 mm. For each of these small squares, we do a ray cast from viewing point towards the center of the small square. If this ray cast hits an occluder before hitting the target, we will mark this small square not visible, otherwise, mark it visible. After doing ray cast for all of the small squares, we count the total visible squares and divide by total number of squares on a single target. We then get our measurement of visible percentage for this target. By comparing the visible percentage of each target, the ideal observer will select the one with higher percentage as closer.

5.2.2 Range cue algorithm

Our model using the range cue is to first identify the occluders in front of the target, then estimate the depth of these occluders using other parallax cues, then find the furthest point of these occluders, and use it as the lower bound of target depth.

To mimic range cue by algorithm, same as above, we first divide the target into a grid of small squares, each the size of 0.5 mm by 0.5 mm. For each square, we do a ray cast from viewing point towards the center of the square. If this ray cast hits an occluder before hitting the target, we record the distance from the hitting point to the camera on depth axis, otherwise we record 0. After doing ray cast for all of the small squares, we find the longest recorded distance, which is the target depth lower bound. Because there is no other depth information available for this observer, we use this lower bound as estimated depth. We then compare the two estimations for two targets and select the closer one.

Interestingly, one can speculate that humans may use the depths of occluders both in front and behind the targets. To investigate this speculation, we did a simple experiment on ourselves with all occluders in front of the targets removed, versus all occluders behind the targets removed. As observers, we felt that when the behind occluders removed, the depth perception was mostly unaffected comparing to the uniform distribution. But when the front occluders were removed, the task became extremely difficult. Following this, we decided to only include front occluders in the calculation of our range cue ideal observer. However it is worth noting that the visible front and behind occluders are obviously different in numbers and we did not have the chance to test on naive subjects.

5.2.3 Long targets

When using the above algorithms for long targets, most of the process stay the same, except that for the left or right end of the target, raycasts will hit the planes on the side of the clutter, see Figure 1.1b. In this case, we will not include their raycast result in the total count, as those point are designed not be visible to human observers.

5.3 Results and Discussion

5.3.1 Results

Using the two ideal observers stated above, we tested under conditions with the combination of long targets and short targets and four clutter distributions. For each ideal observer and each condition, we tested on 20 different ΔZ. Because the data is more interesting on smaller distances, we selected the distances in log units. When $\Delta Z =$ 20 cm , the targets are always outside of the clutter distribution. The observer result is trivial and do not represent the trend of the curve, thus we omitted those results. For each ΔZ we ran 5000 trials, and record the percentage of correct results from ideal observer. Figure 5.1 & 5.2 shows the results.

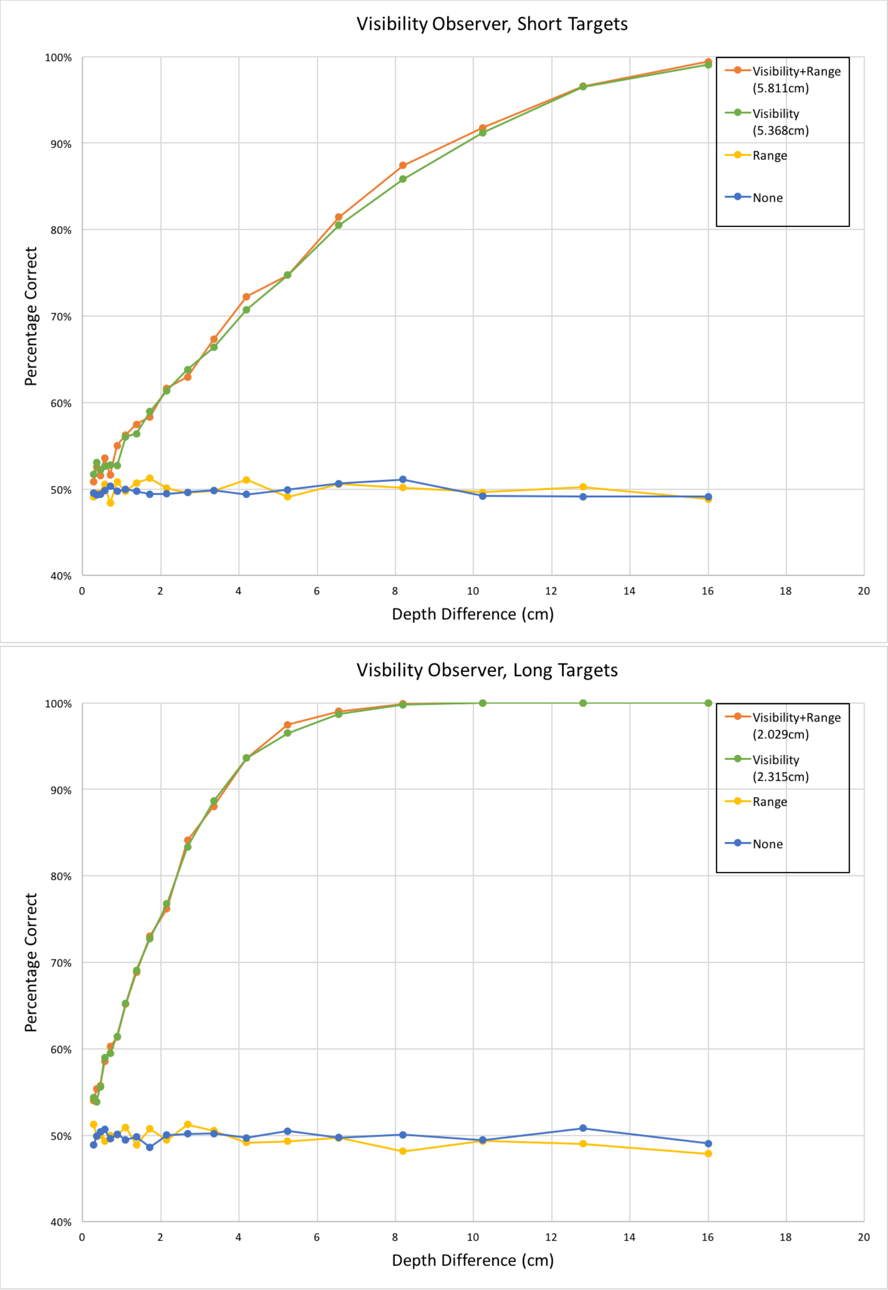


FIGURE 5.1: Results from visibility ideal observer algorithm, each percentage correct is from 5000 trials.

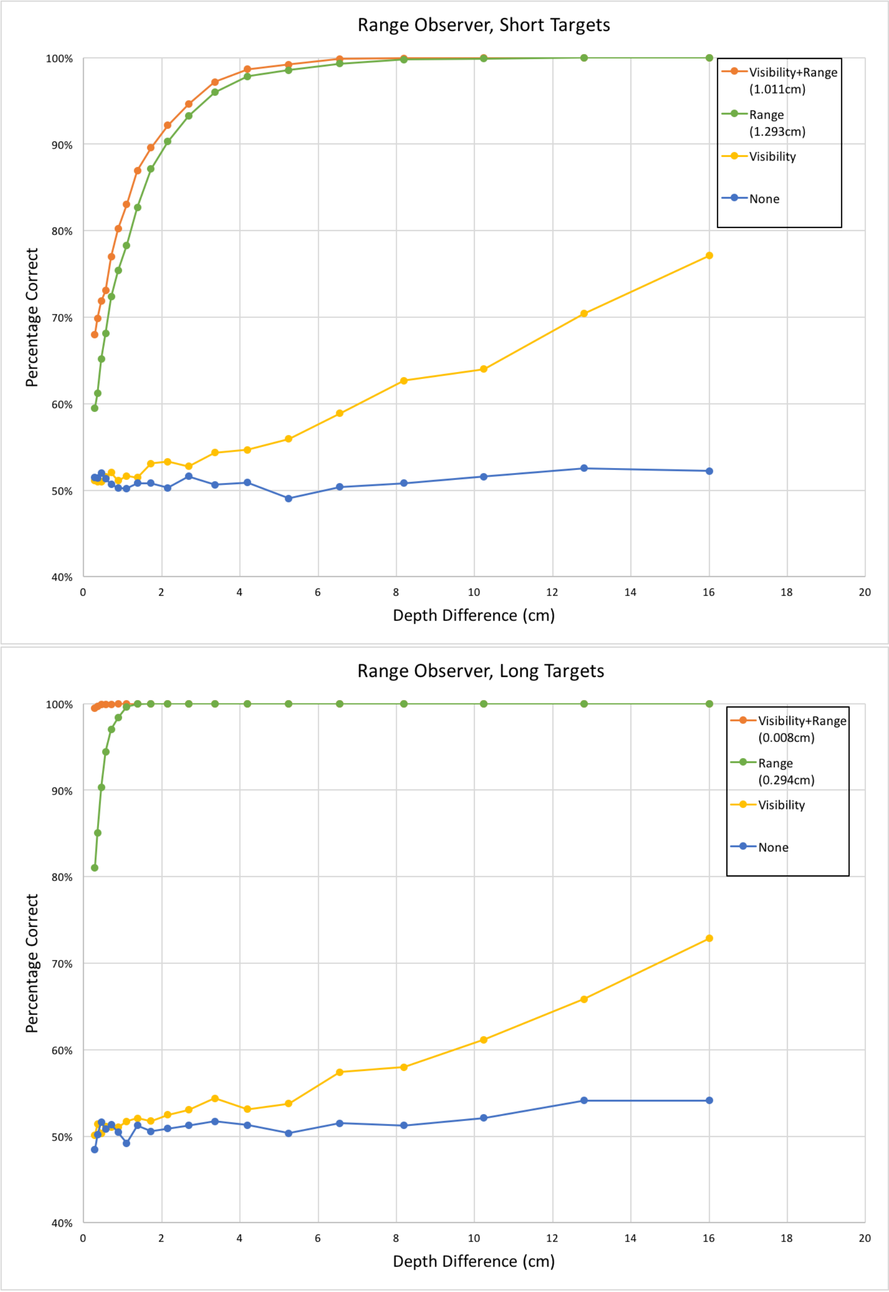


FIGURE 5.2: Results from range ideal observer algorithm, each percentage correct is from 5000 trials

5.3.2 Verifying the design of the experiment

First we used the these ideal observers to run the same experiments as the test subjects, the mean thresholds of 15 runs (same number as number of human test participants) are indicated with brackets in legends of Figure 5.1 & 5.2. It is easy to see that these thresholds roughly correspond to 78% of accuracy, in keeping with our design intend of the experiment.

Secondly, it is easy to observe that whenever the scene contains the cue on which the ideal observer is based, the performance is well above 50% (orange and green lines). When the scene does not contain the right cue, the performance is near random chance (yellow and blue lines). This shows that our design of clutter distribution can effectively isolate either of the occlusion cue.

However, we can also see some above 50% performance using range observers on visibility only distributions, namely the yellow lines in Figure 5.2. This is not our design intent but can be explained if we go back to the distribution illustrations in Figure 3.7. In visibility only distribution, the density of occluders in front of the two targets are different. If we find the furthest occluder among these two, the denser side will be more likely to have a further front occluder, and this will coincidentally give the correct answer. This means our design of this particular distributions cannot eliminate range cue completely. But this effect is very small comparing to other existing cues. For the same clutter distribution, human performances are between 2 cm and 7 cm (Figure 4.1 & 4.2), and for these ΔZ, range ideal observer has at most 60% accuracy rate in visibility cue distributions. Also keep in mind that this range ideal observer has access to the precise depth of every occluder, for which human observers can only estimate. So it is safe to assume that this will not affect human subjects' performance in a significant way, and the performance in visibility distribution is mostly due to the use of visibility cue.

5.3.3 Comparing to the result of test subjects

For the visibility ideal observer, the result is quite similar to that of human subjects when neither of the parallax cues are enabled (grey bars in Figure 4.1 & 4.2). More specifically, for short targets human achieved 5 cm to 7 cm and the visibility ideal observer achieved around 5.5 cm. For long targets humans achieved 3 cm to 4 cm and the ideal observer achieved around 2.2 cm. This indicates that human subjects made nearly full use of available visibility information, and this might look surprising at first glance, but it is still reasonable. This is because, unlike other cues such as binocular disparity, this judgement does not rely on the fine detail of the image, for which ideal observer algorithm holds an advantage. For visibility cue, the number of red pixels is only a statistical cue to depth. Because of the randomness of the clutter, one observation of visibility has very unstable result (recall large error bars from Figure 2.4), making the difference between visible areas of two targets very obvious in most cases. When ideal observer observed a small difference which human eye cannot see, the chance of ideal observer being right is just close to chance. To sum it up, the information of visibility cue relies on the overall probability, not individual observations, that's why human observers can achieve almost as well as ideal observers.

For range ideal observer, however, the performance is very much better than the human counter parts. More specifically, the range ideal observer achieved below 1.5 cm, better than human even in the condition where every cue is available. This is expected as the range ideal observer algorithm is given exact information about nearby occluders' depth, whereas human observers need to estimate the depth from parallax cues.

5.3.4 Stereo ideal observer for visibility cue

As mentioned before, by analysing results from human subjects' performance for long targets, we can see that parallax cues do not improve performance of visibility cue. This is somewhat counter-intuitive, because we expect that with multiple viewpoints, visibility will become less variable and thus there will be more precise predictions. To investigate this issue, we developed a binocular version of visibility the ideal observer. It first calculates the visible percentage of each target from each viewpoint i.e. each eye, then added the result from left eye to result from right eye, then compared the sum. We show the performance of the stereo version versus the original mono version on short targets in Figure 5.3.

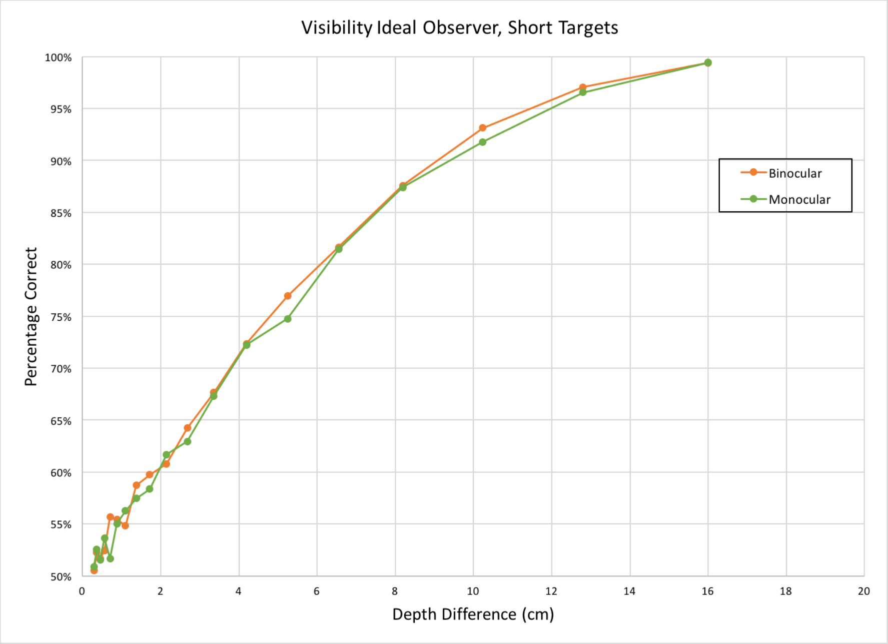


FIGURE 5.3: Performance of binocular and monocular ideal observers, each data point is from 5000 trials

Here we can see a very slim improvement between binocular observer and monocular observer. It shows that there is not much additional information from binocular vision, contrary to our initial thought. Here we will explain why. In stereo, when the observers try to determine the depth, they use the information from both eyes. In our particular setup, the generated scene is fairly far from the viewpoint (60 cm away), and the binocular distance (6.4 cm) is comparably very small. Thus the difference between left and right eye images are very small. Indeed human eyes are capable of picking up these small differences for stereo fusion, but for the visibility cue, the observers are merely using them for the visible area comparison, rather than trying to use the difference to infer depth directly. Thus the small difference means it will not help the accuracy very much (recall large error bars from Figure 2.4). Motion parallax is very similar to binocular vision, just with more viewpoints, but these viewpoints are still very close to each other, thus the above argument still works. From these analysis, we can see it is not surprising that both parallax cues do not improve the visibility cue significantly.

5.3.5 Interaction with perspective cue

To better study the visibility cue by itself, we eliminated the well-known perspective cue for all of our experiments. However, it is still interesting to see how well the visibility observer makes use of perspective cues. We modified our visibility ideal observer, so that it will not just return the visible percentage, but return the raw number of visible pixels of each target to compare, and we used this observer for short targets with and without controlled size. In addition, we tested it on a clutter that removed visibility cue and target size uncontrolled.

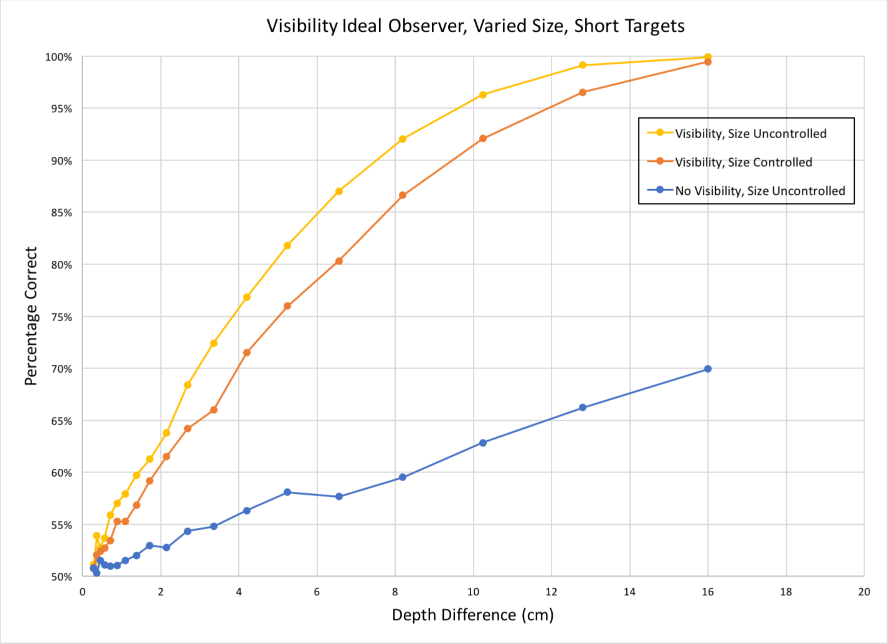


Figure 5.4: Comparing results for different size conditions. Each data point is from 5000 trials

We can see that with target size uncontrolled, the performance improved as much as 10%. If you use this new visibility observer on condition that contains only perspective cue, the performance is still relatively low.

It is interesting to note that there is more than just the visibility difference from perspective cue. Because our targets are regularly shaped rectangles, the observer can actually try to use the ”visible height” or ”visible width” of the target to estimate its size more precisely. To investigate this, we developed another perspective ideal observer. Its algorithm is stated as follow.

To mimic the process of estimate ”visible size”, we first divide the target into a grid of small squares, each has the size of 0.5 mm by 0.5 mm. For each square, we do a ray cast from viewing point towards the center of the square. If this ray cast hits an occluder before hitting the target, we record its y position. After doing ray cast for all of the small squares, we find the top most and bottom most y position, and use their distance as the estimation of the height of the target. We then compare the two estimated heights for two targets and select the one with larger height as closer.

We run this observer on small targets with uncontrolled size, the result is shown in Figure 5.5

As we can see, this observer performed extremely well, because this cue relies heavily on the ability to tell very small differences. For this reason, we believe human vision is not able to make full use of this cue, unlike visibility cue.

Overall, although we did not include perspective cue in our human subjects' experiment, we believe its presence can also improve depth discrimination performance in cluttered scene, either with visibility cue or without. This can be an interesting area to explore in future studies.

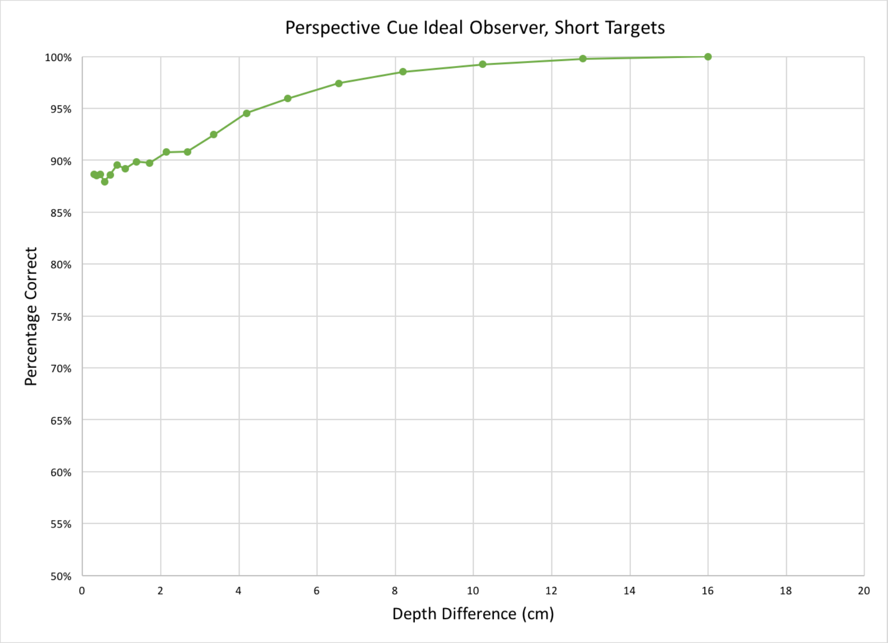


FIGURE 5.5: Result from perspective cue ideal observer, each data point is from 5000 trials

Chapter 6

Conclusions

Our experiments and analysis have provided new insights into depth perception in 3D cluttered scenes, in particular, scenes in which the clutter is dense and so the effects of occlusions cannot be ignored. We have identified two new metric occlusion cues to depth in 3D clutter, namely a visibility cue and a range cue. We have shown how humans combine these depth cues with binocular disparity and motion parallax. One might have expected that 3D clutter simply interferes with depth perception by reducing the information from binocular disparity and motion cues. Our experiments show that the situation is more complicated than that. Occlusions also provide depth information which observers use to discriminate the depths of identifiable targets that are embedded within the 3D clutter. The depth information provided by occlusions does not fully compensate for the loss of information from reduced visibility, but in some situations it comes close.

More generally, 3D cluttered scenes provide a rich and natural but neglected domain for studying depth perception. We have concentrated on how occlusion cues are combined with stereo and motion parallax. But other cues should be examined as well, including perspective and shading. Finally, 3D clutter is common in natural scenes, but there has been little work in vision science to quantify how common it is and what the implications are [6]. We hope that some of the ideas of this thesis could stimulate the community to address these questions.

Bibliography

[1] S. Rezvankhah, “Depth discrimination in cluttered scenes using fishtank virtual reality,” Master’s thesis, McGill University, May 2015.

[2] B. Julesz, “Binocular depth perception of computer-generated patterns,” *Bell System Technical Journal*, vol. 39, no. 5, pp. 1125–1162, 1960.

[3] J. I. Yellott, “Foundations of cyclopean perception,” *Behavioral Science*, vol. 17, no. 3, pp. 310–312, 1972.

[4] B. Rogers and M. Graham, “Motion parallax as an independent cue for depth perception,” *Perception*, vol. 8, no. 2, pp. 125–134, 1979.

[5] B. Rogers and M. Graham, “Similarities between motion parallax and stereopsis in human depth perception,” *Vision Research*, vol. 22, no. 2, pp. 261 – 270, 1982.

[6] M. A. Changizi and S. Shimojo, “X-ray vision and the evolution of forward-facing eyes,” *Journal of Theoretical Biology*, vol. 254, no. 4, pp. 756 – 767, 2008.

[7] M. S. Langer and F. Mannan, “Visibility in three-dimensional cluttered scenes,” *Journal of the Optical Society of America A*, vol. 29, no. 9, pp. 1794–1807, 2012.

[8] K. Nakayama and S. Shimojo, “Da vinci stereopsis: Depth and subjective occluding contours from unpaired image points,” *Vision Research*, vol. 30, no. 11, pp. 1811–1825, 1990.

[9] J. M. Harris and L. M. Wilcox, "The role of monocularly visible regions in depth and surface perception," *Vision Research*, vol. 49, no. 22, pp. 2666–2685, 2009.

[10] R. A. Akerstrom and J. T. Todd, "The perception of stereoscopic transparency," *Perception & Psychophysics*, vol. 44, no. 5, pp. 421–432, 1988.

[11] I. Tsirlin, R. S. Allison, and L. M. Wilcox, "Stereoscopic transparency: Constraints on the perception of multiple surfaces," *Journal of Vision*, vol. 8, no. 5, p. 5, 2008.

[12] G. J. Andersen, "Perception of three-dimensional structure from optic flow without locally smooth velocity.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, no. 2, p. 363, 1989.

[13] R. van Ee and B. L. Anderson, "Motion direction, speed and orientation in binocular matching," *Nature*, vol. 410, no. 6829, pp. 690–694, 2001.

[14] J. M. Harris, "Volume perception: Disparity extraction and depth representation in complex three-dimensional environments," *Journal of Vision*, vol. 14, no. 12, p. 11, 2014.

[15] I. E. Sutherland, "A head-mounted three dimensional display," in *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, pp. 757–764, ACM, 1968.

[16] P. R. Desai, P. N. Desai, K. D. Ajmera, and K. Mehta, "A review paper on oculus rift-a virtual reality headset," *arXiv preprint arXiv:1408.1173*, 2014.

[17] K. W. Arthur, K. S. Booth, and C. Ware, "Evaluating 3d task performance for fish tank virtual worlds," *ACM Transactions on Information Systems (TOIS)*, vol. 11, no. 3, pp. 239–265, 1993.

[18] S. M. LaValle, A. Yershova, M. Katsev, and M. Antonov, "Head tracking for the oculus rift," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pp. 187–194, IEEE, 2014.

[19] C. Blakemore, "The range and scope of binocular depth discrimination in man," *The Journal of Physiology*, vol. 211, no. 3, p. 599, 1970.

[20] M. A. Garcıa-Pérez, "Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties," *Vision Research*, vol. 38, no. 12, pp. 1861–1881, 1998.

[21] E. B. Johnston, B. G. Cumming, and M. S. Landy, "Integration of stereopsis and motion shape cues," *Vision Research*, vol. 34, no. 17, pp. 2259–2275, 1994.

[22] M. F. Bradshaw and B. J. Rogers, "The interaction of binocular disparity and motion parallax in the computation of depth," *Vision Research*, vol. 36, no. 21, pp. 3457–3468, 1996.

[23] R. L. Sollenberger and P. Milgram, "Effects of stereoscopic and rotational displays in a three-dimensional path-tracing task," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 35, no. 3, pp. 483–499, 1993.

[24] L. M. Wilcox and R. S. Allison, "Coarse-fine dichotomies in human stereopsis," *Vision Research*, vol. 49, no. 22, pp. 2653–2665, 2009.