Copyright

by

Brian Clark McCann

2015

The Dissertation Committee for Brian Clark McCann Certifies that this is the approved version of the following dissertation:

Naturalistic Depth Perception

Committee:

Wilson S. Geisler, Co-Supervisor

Mary M. Hayhoe, Co-Supervisor

Lawrence K. Cormack

Alexander C. Huk

Alan C. Bovik

Naturalistic Depth Perception

by

Brian Clark McCann, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2015

Dedication

This work is in memory of David C. Knill; you taught me true perspective.

Naturalistic Depth Perception

Brian Clark McCann, PhD

The University of Texas at Austin, 2015

Co-Supervisor: Wilson S. Geisler

Co-Supervisor: Mary M. Hayhoe

Making inferences about the 3-dimensional spatial structure of natural scenes is a critical visual function. While spatial discrimination both in depth and on the image plane has been well characterized for simple stimuli, little is known about our ability to discriminate depth in natural scenes, particularly at far distances. To begin filling in this gap we: (i) developed a database of 80 stereoscopic images paired with the corresponding measured distance information, (ii) used these scenes as psychophysical stimuli and measured near-far discrimination acuity in 4 observers as a function of distance and the visual angle separating the targets, (iii) made additional measurements under patched-eye (monocular) viewing conditions to evaluate the importance of binocular vision in depth discrimination as a function of viewing geometries.

We find that binocular thresholds are roughly a constant Weber fraction of the distance for absolute distances ranging from 4 to 28 meters. Further, measured thresholds were around 1% for small separations, and increased to 4% for stimuli separated by 10 deg. Thus, the ability to discriminate depth in natural scenes is very good out to considerable distances. To investigate the basis of this discrimination ability, monocular thresholds were measured. We found that monocular thresholds were

elevated for distances less than 15 meters, but were comparable to binocular thresholds for greater distances.

Accurate depth perception depends on combining (fusing) multiple sources of sensory information. Thus binocular thresholds probably involve fusing separate monocular and disparity-derived estimates. Under the assumption of Gaussian distributed independent estimates, Bayes rule provides a simple reliability-weighted summation model of cue combination. Using disparity threshold measurements by Blakemore (1970), and the current monocular thresholds, parameter-free predictions were generated for the current binocular thresholds. These predictions were in broad agreement with the data, suggesting that the disparity and monocular cues are separable and combined optimally in natural scenes.

Table of Contents

List of Figures

Chapter 1: Background

1.1 – Overview

Motivation:

Our sensory systems have evolved under the constraints imposed by natural tasks and natural environments. Inferring the 3d geometry of a scene is presumed to have been a perceptual function critical to the survival of humans and other animals. Therefore, it is plausible that there has been considerable pressure on the organizations of the brain and sensory systems to precisely infer 3d geometry in the context of natural environments (Geisler and Diehl 2002).

Major Contributions:

Oddly enough the basic spatial acuity of human depth perception has not been adequately characterized. The reasons for this failing are complex, and are discussed in more detail over the following sections. The current work makes three major contributions: (i) A database is constructed that relates the 3d structure of natural scenes to the sensory stimuli they generate, (ii) basic spatial acuity for depth is characterized for these scenes, and (iii) the relative importance of binocular and monocular cues are assessed over a wide range of viewing conditions.

Natural Scenes:

Natural scenes are often (incorrectly) thought of as just pictures. In general, the 'scene' does not refer to a picture; it refers to the environment itself. Thus, pictures are better thought of as images of scenes. That is, the sensory stimulus generated by the natural environment via some generative process.

Here, the scene properties of interest are its 3d structure, and the resulting stereoscopic image. I stereoscopically imaged a large set of scenes using a camera with well-characterized spatial and chromatic imaging properties. I collected the 3d structure of these same scenes using a scanning laser range finder that produces 'range images.' I then determined (with appropriate calibration) the mapping between the 3d positions of the range pixels, and pixels in the stereoscopic photographs. This dataset is of inherent value for understanding not only the physical structure of scenes, but also their relationship with associated visual projections. Care was taken to ensure the measurements are of a high technical quality. Database construction methods are detailed in Chapter 2.

Acuity Measurements:

What do we mean by depth acuity? It is known that performance in depth discrimination tasks depends large number of factors. Some have attempted to understand depth acuity by repeating measurements over a small set of carefully designed stimuli. Usually, these stimuli are presumed to isolate a single component of more general depth perception models. For example, acuity for the component cue of binocular disparity has been well characterized (Blakemore 1970). Perhaps by integrating experimental knowledge of this sort it will ultimately be possible to make accurate predictions for arbitrary stimuli, including arbitrary natural scenes. While efforts to develop models from experiments with laboratory stimuli have been fruitful, there is still no model capable of making such predictions.

An alternative approach to studying depth acuity is to directly measure performance in real scenes. For example, depth estimation studies have been run outdoors (Gibson and Bergman 1954), or in a long corridor (Holway and Boring 1941). These

studies have the advantage that they are of direct relevance to the stimuli of interest, i.e. natural environments. Using stimuli from varying 'real world' conditions allows for stable measures of average performance appropriate for stimulus naïve benchmarks. However, laboratory displays have practical advantages, for examples better stimulus control, and they reduce the effort necessary to run a large number of trials.

The current work attempts to combine the advantages of each of these approaches by bringing natural scenes into the lab. Measured stimuli are reproduced on a laboratory projection screen with the goal of exactly replicating the stereoscopic view of the scene through a window. The specific acuity task under consideration is the near-far distance judgment of two arbitrary points corresponding to points on surfaces in natural scenes. Since the ground-truth positions of the imaged surfaces are known, observer judgments can be evaluated objectively. Methodological details regarding psychophysical tasks are discussed in Chapter 3, while results of the psychophysical experiments are discussed in Chapter 4.

Cue Combination Framework:

The depth perception literature has largely been concerned with characterizing and understanding the different sources of depth information (cues) available to the visual system. The relationship between binocular and monocular cues to depth has been of particular interest. However, the viewing conditions under which the binocular cues (most notably disparity) are relevant to performance have still not been fully characterized. Geometry suggests that disparity will be less useful at large distances. However, the quantitative extent to which this degradation will impact depth discrimination in natural scenes is unknown.

One major impediment to understanding the relative importance of binocular vision at large distances has been our lack of knowledge about monocular depth cues. Neither the importance nor the prevalence of monocular depth cues in images is known. Consequently, the importance of binocular cues *relative to* monocular cues cannot be known.

The current work attempts to directly address this gap in knowledge. Until now, it was not possible to measure the precision of monocular depth estimates in real scenes and at large distances. As mentioned above, it has been logistically difficult to run experiments at large distances, especially over many trials and with access to precise ground-truth data. Our psychophysical approach makes this possible. Here we can directly measure discrimination performance with one eye, and compare it to discrimination performance with two eyes for a given distance and angular separation.

Importantly, some of the computational modeling concepts the have been applied in experiments with artifactual stimuli are still applicable in the present case. For example, the concepts of Bayesian cue combination, which have been applied to e.g. random dot stereograms (Ernst & Banks 2002), can also be applied here using natural scenes (see later and section 1.2 for review of models of depth perception). The value of the present approach derives from marginalizing across the variability inherent to naturalistic stimuli while still allowing rigorous analysis and modeling.

1.2 –DEPTH PERCEPTION: THEORIES

Guiding Principles:

The optic nerve is the brain's only visual access to the world, and hence estimations about the world can only be made with information conveyed by the visual stimulus as represented in the images formed on the retinas. Conversely, survival depends on reasoning about the world, not the stimulus at the eye. Perception is the process of associating stimulus properties, such as luminance, color and disparity, with environmental properties, such as the distance and material compositions of surfaces (Adelson 1995). One useful way to conceptualize problem of perception is that it constructs an 'internal model' (Figure 1), which is a parameterized representation of the environment, obtained in some way via neural computations (Rao 1999). Parameters of interest are usually part of the 'hidden state' of the environment. Thus, they must be estimated from the 'visible state' (sensory stimulus).

Depth perception is the process of associating a physical stimulus with a spatial structure (the difference in distance between two points). There are many sources of information (cues) available for depth estimation and much of the theory of depth perception concerns describing the cues and how they are used by the visual system.

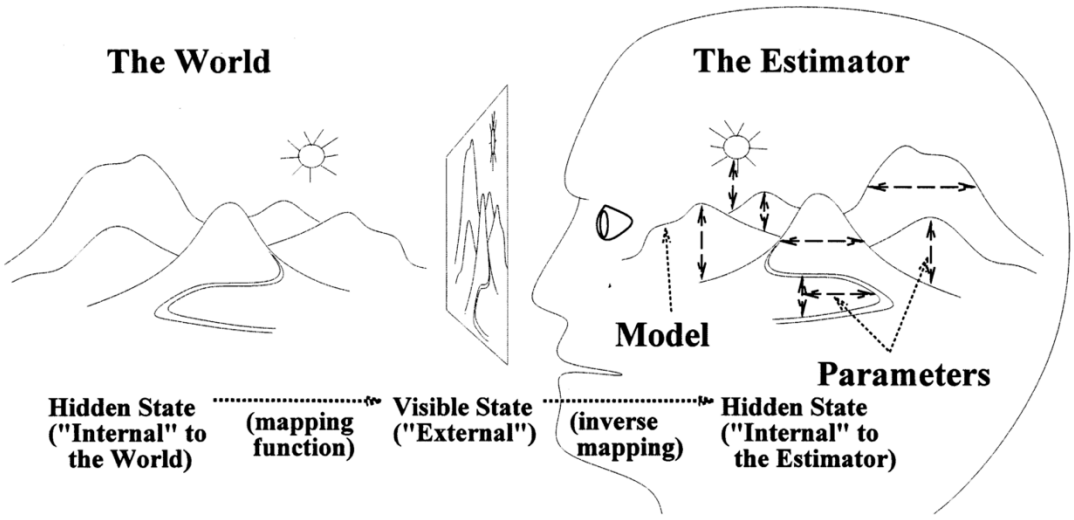

Figure 1 - Schematic Representation of the Internal Model

'The Estimator' creates a parameterized representation of the world, called the 'Internal Model.' These parameters reflect properties of the 'hidden state' of the world, e.g. distance relationships between objects. Therefore, the value of these parameters must be estimated (via some inverse mapping) from the 'visible state,' i.e. the visual stimulus. The hidden state and its associate mapping function constitute the generative model of the stimulus, (from Rao 1999 and based on O'Reilly 1996).

Depth Cues:

Depth cues are sources of information about depth. The term is intentionally vague; cues can have many physical sources. With regard to this work two classes of cues are of interest, specifically stereoscopic (or binocular) cues to depth and pictorial (or monocular) cues to depth. There are other very important depth cues, but they were not the primary cues under study. See *Perceiving in Depth* (Howard and Rogers 2012) for a more exhaustive review of the historical depth perception literature.

Depth Cues - Historical Context:

Giotto Di Bondone, an artist from the 14th century, is credited with being the first to have treated a painting like a realistic view through a window. What he, and subsequent renaissance artists had begun realizing is that the illusion of depth in a painting can be made more realistic by following the rules of linear perspective (Edgerton 1991). Later, during the 19th century, Charles Wheatstone (Wheatstone 1838) first demonstrated how two planar images presented to the two eyes cause an observer to experience an even more realistic depth percept. For at least two centuries, it has been widely known that the brain is sensitive to many depth cues. The particular cues of interest here are exactly those you would find in paintings and stereoscopes. That is, static monocular cues to depth like perspective, occlusion and lighting cues, as well as static stereoscopic cues to depth like binocular disparity.

Depth Cues – Disparity Primer:

Binocular disparity is the most comprehensively studied cue to depth. A disparity results from the horizontal displacement of our eyes. Disparity is closely related to parallax in that both cues are generated by a change in viewing position. In order to get a common sense understanding of disparity, cover one eye and move your head to the side such that your open eye is now in your closed eye's original position. The motion you observe is known as 'motion parallax.' However, this same difference in the image is used by the binocular disparity system without the need to integrate over time. You may note that if you are fixating at a distant point, the relative translation of nearby objects tend to be greater than the relative translation of distant objects. This same principle holds for stereopsis.

Binocular disparity specifically refers to the angular difference in the projected location of an object between the two retinal projections (see Figure 2a). An easy way to calculate disparity in complicated situations is to consider the convergence angles of the eyes necessary to fixate the targets (known as vergence demand). The difference in these angles is the relative horizontal disparity. Absolute horizontal disparity, on the other hand, depends on the point of fixation (i.e. the eyes' convergence angle), not just the vergence demands of the targets.

Subtle complications in stereo geometry arise quickly. Gerhard Vieth and Johannes Müller are credited with formalizing the horizontal horopter, the locus of corresponding points in the two eyes (see Figure 2b). The so-called Vieth-Müller circle traces a perfect circle through the nodal points of two eyes, and the point of convergence (Müller 1826). Therefore, the relationship between depth and horizontal disparity is not isometric, i.e. zero disparity does not imply zero depth. However, a point in 3-space can be fully specified given a disparity in addition to other ancillary information, e.g., the convergence, azimuthal (version), and elevation angles of fixation, as well as the viewer's position and inter-pupillary distance (IPD).

People are extraordinarily sensitive to disparity. As far back as World War I thresholds for disparity detectability were already being measured as low as two seconds of arc (Howard 1919). Such a low threshold is especially remarkable considering that the smallest photoreceptors (pixels) in the human eye subtend roughly thirty seconds of arc. Thus disparity is a so-called 'hyper-acuity.'

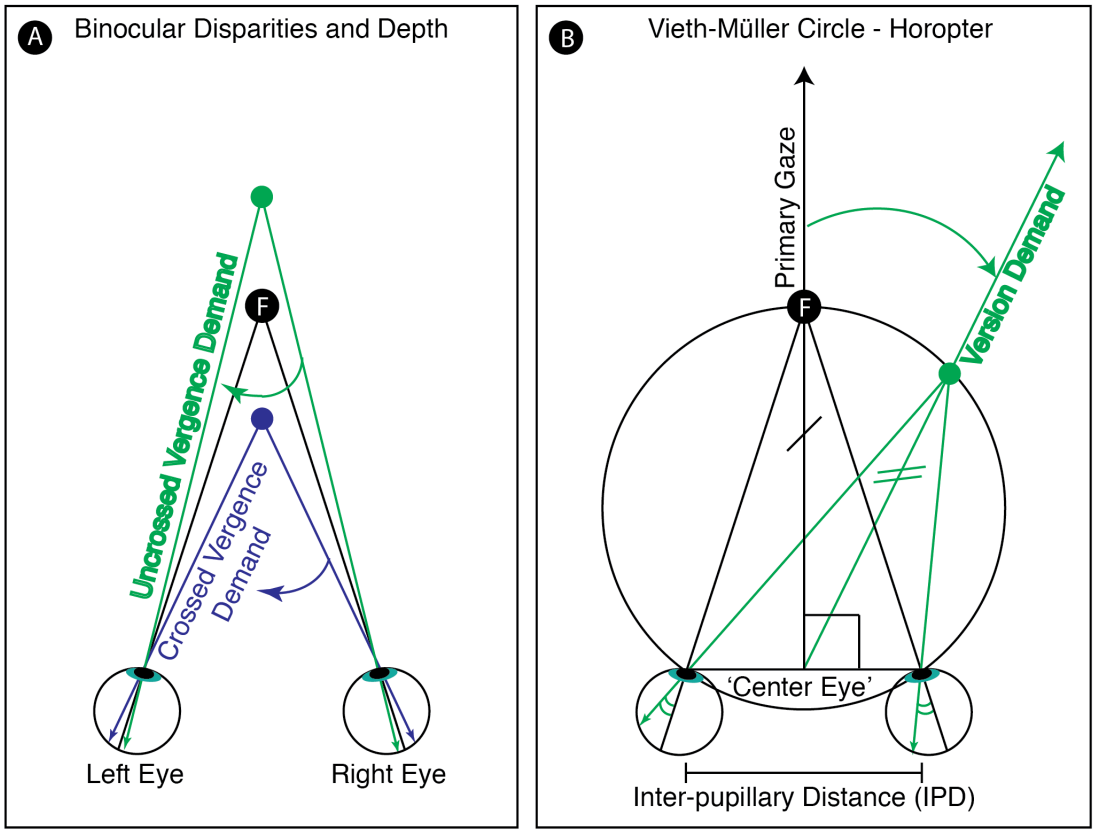

Figure 2 – Binocular Disparity Geometry

In both images, the two eyes are drawn converged on the fixation target (F). The black lines through the eyes thus indicate the optical axis of the eye. (A) A schematic of the relationship between changing vergence demand, depth, and relative horizontal disparity. It can be seen from the image that even for the same magnitude of depth, crossed disparities are larger than uncrossed disparities. (B) A schematic of the Vieth-Müller circle showing the impact of version demand. 'Primary Gaze' is the axis orthogonal to the inter-ocular axis as well as the ground-plane (not depicted). Version demand is the magnitude of the azimuthal eye movement necessary to fixate the target. Notice that with

non-zero version, a target can have zero disparity (double hashed angles), and non-zero depth (unequal length hashed lines). The inter-pupillary distance (IPD) is also depicted. Average IPDs are around 65mm for males.

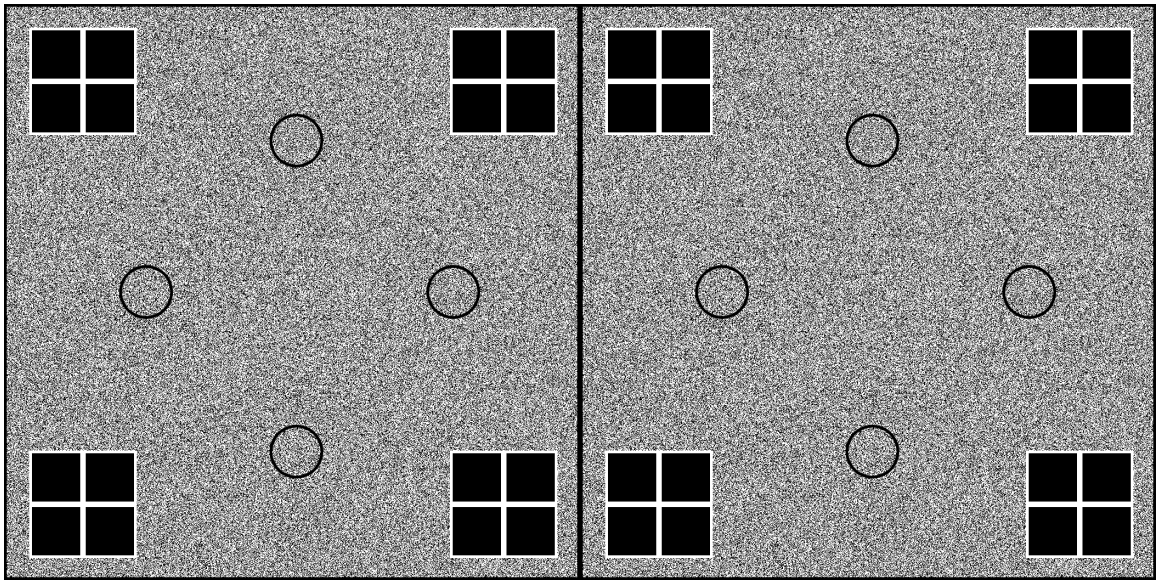

Figure 3 – A Random-dot Stereogram

Cross your eyes to fuse the registration marks in the two images. A depth-defined shape emerges in the middle of the 'cyclopean' view. Stimuli like this one are thought to isolate the binocular disparity cue to depth.

The Gestalt psychologists were of the opinion that the monocular image needed to be grouped into objects for binocular disparity to be a useful cue to depth. Béla Julesz (Julesz 1960) showed that cleverly arranged random-dots could be stereoscopically fused to create a shape defined by illusory depth. Importantly, the shape is ill defined in both monocular images. Random-dot stereograms (see Figure 3) provide the canonical

example of an isolated depth cue. The only information in the image defining the cyclopean shape is binocular disparity. Therefore, disparity must be the information used by the observer, and it can be used independently of any pictorial cues.

Depth Cues – Painters' Cues:

The increased realism of renaissance art came from the geometrically correct use of pictorial cues to depth. All pictorial cues to depth depend entirely on assumptions about the world. Linear perspective rests on the assumption that converging lines are more likely parallel lines receding in depth. Relative size cues depend on the assumption that two objects are in actuality the same size. Texture density cues depend on the assumption that these regularly sized texture elements are equally spaced on surfaces. Indeed, many monocular cues derive their information content from the geometry of perspective projection. Under perspective projection, objects more distant in the scene appear smaller in the image. Thus knowledge (relative or absolute) about the size or shape of an object can be informative about its position and orientation in 3d. However, even perspective cues are only informative in conjunction with some *a priori* knowledge.

Other geometric regularities in the world contribute to monocular depth perception as well. Occlusions are more likely than concave cutouts viewed from the perfect angle. Lights tend to be above, suggesting shaded surfaces are either raised or depressed. In some cases our familiarity with a particular class of objects provide very strong evidence of a particular interpretation e.g. walls meet at right angles or faces are more likely than hollow masks. Presumably we believe in these structural assumptions about the world only insofar as they hold in our broader experience. On average people are assumed to be unbiased depth estimators. However, strong mechanisms can be

exploited by clever stimulus design to create illusory depths (biased depth perception). Typically, depth illusions depend on incorrect structural assumptions about the stimulus.

Depth Cues – Consistency and Conflict:

Depth cues are commonly considered to be (i) consistent with a particular 3d interpretation, (ii) in conflict with a particular 3d interpretation, or (iii) uninformative about a particular 3d interpretation. Cues that are consistent with wildly differing 3d interpretations are thus said to conflict. Weak cues are typically not considered consistent or in conflict with anything, rather they are uninformative. The difference between a (relevant) conflicting cue and an (irrelevant) uninformative cue is worth considering when designing experiments.

Renaissance paintings provide an illustrative example of wildly conflicting cues. Pictorial cues to depth, for example shape cues, size cues, texture cues, and lighting cues, are all plausibly consistent with the depicted non-planar interpretation of the scene geometry. Other depth cues are in conflict with the pictorial interpretation. For examples the pattern of defocus, the motion parallax as well as the pattern of binocular disparities are all strongly consistent with the veridical planar interpretation of the canvas. Considerable effort has been made to formalizing models of rational data fusion. These models of depth perception are discussed in more detail under 'Combining Cues.'

It is worth addressing the difference between conflict and irrelevance. Continuing with the painting example, binocularly viewing the flat image introduces overwhelming visual evidence that the image is, indeed, flat. The pattern of disparities produced by the painting would be extraordinarily unlikely for any interpretation of the physical scene inconsistent with a planar canvas. Alternatively, if the painting were viewed with a

patched or damaged eye, the painting would be more difficult to discriminate from an actual view of the depicted physical scene.

In the former case, responses from the second eye are clearly in conflict with the depicted scene. In the latter case, the brain plausibly ignores responses from the second eye because they are relatively uninformative about the scene's physical interpretation. Sometimes we close an eye. It has no relevance on the structure of the outside world. Therefore, the responses (or lack there of) from the second eye are not conflicting, rather they are just so uninformative that they are irrelevant to the adopted scene interpretation. In the painting example, other cues would still be informative, notably parallax and defocus. However, with a stable head and enough viewing distance these cues could be made uninformative as well. Interestingly, humans can make judgments of 3d structure in binocularly viewed paintings, even if they are simultaneously perceived as a painting on a flat canvas. However, when comparing the relative influence of various cues it is good practice to assure that there is no cues are providing extremely biased (conflicting) information.

Combining Cues:

In the case of conflicting cues, it is not *a priori* clear what sensory interpretation should be adopted. Most models of depth perception formalize depth cues as distinct, randomly distributed depth estimates that are subsequently combined by some rule of combination. As Larry Maloney and Mike Landy point out (1989) people are almost certainly 'robust' fusers. That is, when cues suggest wildly conflicting interpretations, the less reliable cue is simply ignored. There is no real need to test extreme cases. We can all look at a painting and know that it is painting (although we also unambiguously agree on the depth relations depicted by the painting).

However, there are more realistic situations where the combination rule is interesting to study in quantitative detail. Models vary in their assumptions, but one common approach is to assume that cues produce depth estimates that are independent, unbiased and Gaussian distributed. Under these assumptions, the optimal combination rule is to linearly weight the various depth estimates inversely proportional to the variability associated with those estimates. Refer to *Data Fusion for Sensory Information Processing Systems* (Clark and Yuille 1990), or *Perception as Bayesian Inference* (Knill and Richards 1996) for review.

These distributional assumptions are unlikely to hold. However, it is a good starting place, and models have already been developed that allow some of the assumptions to be relaxed. For example, the optimal combination rule has been derived for correlated estimates (Oruç, Maloney et al. 2003) allowing for a more complete representation of the jointly distributed cues. Optimal combination rules have been shown to be consistent with performance in a number of experiments covered in greater detail throughout section 1.3.

1.3 –Depth Perception: Relevant Findings

Historical Cue Manipulation:

There has been experimental evidence that cues are relevant to the perception of real 3D structure for quite some time. Holway and Boring (1941) provide perhaps the closest analog to the current work; however they were interested in the perception of object size. Recognizing that the perception of size is integrally linked to the perception of distance they designed an experiment in a long hallway to test their claims. As expected, subjects correctly interpreted the size of an abstract object at the end of the

hallway as compared to a similarly shaped nearby object. That is, they demonstrated size constancy.

In other conditions they then systematically obscured references to the length of the hallway by removing the illumination, and obscuring the walls with black felt. The more comprehensively they removed the cues to depth, the more subjects reverted to a visual angle rule when comparing object sizes. That is, the difference in distance between the two objects was no longer relevant to the observer's percept of the object's physical size without other visual cues to the length of the hallway. Another way to interpret these data is that the observer's estimate of the object's distance down the hallway was so unreliable that the only cue that could be relied on is the presumed similarity in size with a comparable object. That is, the relative size cue was isolated.

Gogel was concerned that these studies all measured non-scalar (relative) percepts rather than metric quantities. His concern was that in reduced cue experiments where two objects at different distances are seen to be at the same distance it is not at all clear what distance they were both perceived to be at. He called this bias towards a particular percept under reduced cue conditions a specific distance tendency (Gogel 1972). Under the Bayesian framework it can be thought of as evidence that observers default to some prior estimate of distance in the absence of disambiguating information. Gogel took this as evidence that we must be careful when designing cue-deprived stimuli. In the Holway and Boring experiment depriving the observers of cues that the objects were at different distances effectively served as evidence that they were in fact at the same distance. Whether that perceived distance was close enough to be held in your hand or further than could be reached in a day's hike is completely unknown to the experimenters.

Rather than completely isolating cues even rich artificial stimuli can be designed to produce biased percepts. Adelbert Ames' spinning trapezoidal window and trapezoidal room were among the first demonstrations of the importance of environmental assumptions in the perception of 3D shape (Ames Jr 1951). In particular that observers have a tendency to assume that indoor spaces are dominated by right angles and that textures (e.g. tile floors) are homogenous. When viewed from a specific 'accidental viewpoint' Ames' trapezoidal room appears to be a regular room. In fact, the percept of a square room is so strong that observers are willing to adopt the interpretation that people walking from one corner to the other are dramatically changing in size as they do so. This demonstrates the remarkable strength of structural assumptions about the world; they can cause us to perceive radical changes in size when we know no such change could occur. However, the illusion is brittle to changes in viewpoint or binocular viewing. Our depth mechanisms work exactly because these deceptive stimuli are impossibly unlikely until they are experimentally imposed.

Specificity of Cue Combination:

As Maloney and Landy point out (1989) it is possible to devise experiments to specifically test the validity of the combination rule in their model. The typical test involves running three experiments to obtain behavioral estimates of the parameters in their model. In the first two experiments each of the two cues under study are isolated and the psychophysical reliability of the cue is assessed. In the third experiment the two cues are combined so that their combined reliability can be compared to predictions. Additionally, small conflicts can be added to the cues so that the specific weight being given to each cue can be measured in the psychophysical bias. It has been repeatedly shown experimentally that pairs of isolated cues are combined in the way that would be

expected by the cue combination model (see Landy, Maloney et al. 1995 for an early review). The stimuli in these experiments were carefully designed to expose precise mechanistic details that continue to hold in a large variety of contexts. As an example of the breadth of these findings, optimal combination has been shown for multimodal combination (Ernst and Banks 2002), in addition to stimuli more relevant to this work that combined binocular and monocular cues (Knill and Saunders 2003).

It is not known how to apply these results to natural scenes. Whatever the brain does is likely reasonable in a wide variety of contexts given these findings. One relatively unexplored area is how to predict cue reweighting as information content changes from location to location or experimental stimulus to experimental stimulus. It would be nice to know how plastic the mechanism is. That is, under what situations can the brain adapt to regularities in a given context. David Knill showed that over the course of an experiment, observers can learn to 'switch models' pertaining to the relevance of isotropy assumptions for his stimuli (Knill 2007). It is not yet clear what the constraints on the brain's model switching mechanisms are, but the results indicated that observers dynamically update their weights appropriately. It is plausible that observers have a limited ability to do so. Most of these cue combination experiments are rely on a single class of stimulus (e.g. texture or shape). It is therefore difficult to predict how their results generalize to a situation where important properties of the stimulus vary from trial-to-trial, a plausible occurrence from location to location in natural scenes. See section 1.4 'My Approach' for more discussion of the importance of trial-to-trial variability on stable measurements of baseline acuity.

Performance in Real Scenes:

Some experiments have used real scenes as stimuli, however real scenes introduce practical difficulties. It is difficult to run the large number of trials necessary for a two alternative forced choice (2AFC) experiment. Instead of being inferred from a large number of fine discriminations, estimates are directly reported either by metric report, or the assignment of a comparison such as in Holway & Boring's study. It has been shown that outdoor estimates of distance are relatively accurate on average if the observers have been trained with the appropriate scale (Gibson and Bergman 1954, Gibson, Bergman et al. 1954). However, individuals' judgments tend to erratic, calling into question the relevance of the accuracy of between-subject averages (Fine and Kobrick 1983).

In general, estimation designs have been shown to be dependent on the reporting methodology used, e.g. verbal report, pointing, or open loop walking (Loomis, da Silva et al. 1996, Fukusima, Loomis et al. 1997, Philbeck and Loomis 1997). This has also been a problem for the paddle boards used in slant estimation (Durgin, Hajnal et al. 2010). Generally, naturalistic motoric reporting is the most veridical; probably because it is consistent with the way sensory signals are used normally as part of the sensory-motor loop (Durgin and Li 2011). However, there is still the problem that reporting noise substantially corrupts your only access to the variability associated with the sensory estimate.

Discrimination designs are better; 2AFC tasks account for reporting noise fairly well. Unfortunately, only a few have been performed in real scenes. There is some converging evidence of a perceptual compression of the foreshortened dimension as measured by discriminating lengths of physical objects in and out of the plane (Loomis, Da Silva et al. 1992, Loomis and Philbeck 1999) although those studies rely on repeating

measurements on the same target (i.e. a stick) against the same background (i.e. the ground). It is plausible that performance would vary for other targets as it would from location to location, but this could not be tested because of the practicalities of working with real scenes.

Disparity discriminations have been measured for real distances as well. Acuity for real physical disparities have been discriminated at large physical distances, although under reduced-cue conditions similar to those of Holway & Boring (Allison, Gillam et al. 2009, Palmisano, Gillam et al. 2010). The contribution of monocular information was intentionally minimized to cleanly measure threshold for disparity. There has been no real attempt at measuring depth discrimination thresholds for varying targets in varying scenes as would be expected in natural viewing.

There is some evidence that binocular deprivation (patching an eye) impacts walking performance. Further, the timing of the deployment of gaze while navigating obstacles (Hayhoe, Gillam et al. 2009). Raising the foot higher above the obstacles suggests a strategy consistent with mitigating increased risk caused by sensory uncertainty. However, a rigorous mapping to the quantitative sensory reliabilities is not possible from this sort of measure.

What we would like are rough numbers to build into the model of uncertainty about distance in the real world. A way to know on average how precise our distance estimates will be. Since this has never been measured in natural scenes (until now) we can only base our predictions on the results of laboratory experiments using artificial stimuli that have a tenuous relationship to the real world; particularly to estimating depths at distances beyond those the observer can reach.

Stereoscopic Acuity – Relevant Experimental Findings:

Under the prevailing cue combination model, the importance of binocular information is determined by the relative reliability of stereoscopic and monocular cues. The crux of this work is that we frankly have no idea what that relative reliability is. Nobody has measured it. Historically measures of acuity for binocular disparity were taken as surrogates for 'depth acuity.' Stereo-acuity measurements date back to Helmholtz. Today, normal disparity detection thresholds are thought to be approximately 2 seconds of arc in the fovea, first shown by the US Air Force when testing pilots during World War I (Howard 1919).

On the other hand, performance varies depending on the viewing conditions. Numerous studies have shown that estimates degrade rapidly as targets are moved away from the horopter, with crossed pedestal disparities being worse than uncrossed pedestal disparities (Westheimer and McKee 1977). It is not clear exactly how these factors would impact depth perception with fixation uncontrolled. In general, the magnitude of disparities will depend on the choice of fixation. However, given that performance is worse at high eccentricities (Blakemore 1970), it is likely that discriminating depths of points with larger spatial separations will show elevated thresholds. For reference, detection thresholds for disparity increase to 30 seconds or so at 5 degrees of retinal angle separation in the scene.

There is a widely held belief that disparity will only be useful within interaction space because small disparities correspond to large depths at large distances (Palmer 1999). Using primary gaze (illustrated in figure 2), one can easily see that the vergence demand for the object at infinite distance is zero. This viewing arrangement is called parallel convergence.

Convergence may as well be zero for objects at near infinite distances (e.g. stars, horizons, mountains in the distance). In fact, assuming the standard male average IPD of 65mm (Gordon, Walker et al. 1989) a convergence angle of one arc second would imply a fixation distance of 13.4km. A convergence angle of two arc seconds implies a fixation distance of 6.7km. For comparison, the convergence distance implied by a 4-degree convergence angle would be around 1m, approximately arms length. At those fixation distances, a (threshold) change in vergence demand of one arc second implies a depth more appropriately measured in mm. Thus, the potency of the disparity depth signal degrades dramatically with the absolute distance of fixation.

That said, Liu (Liu, Bovik et al. 2008) measured the presence of super-threshold disparities in real scenes and found they were more common than previously thought. There were even some presumably detectable uncrossed disparities for fixations greater than 15m (although very few). It is plausible that stereopsis is relevant at distances much greater than previously thought. It all depends on whether or not the variability associated with the disparity cue is large relative to the other available cues.

The quality of disparity information will depend on the stimulus itself as well. Proper fusion depends on proper correspondence between the two eyes' images. Binocular images could have multiple ways in which they could be aligned; for example, repeating stimuli like gratings could be matched along any cycle. Horizontal gratings, on the other hand, have no specific horizontal disparity, but may still imply a vertical disparity. Mismatches in luminance yield poor fusion. Small changes in the interocular correlation are fairly detectable assuming super-threshold contrasts (Cormack, Stevenson et al. 1991) and sufficient dot density (Cormack, Stevenson et al. 1997). Glennerster and McKee showed that stereo acuity for two vertical lines is better when the lines are

superimposed on a reference plane of dots (Glennerster and McKee 1999). Real scenes are a mix of relatively uniform areas, and areas of increased feature density and depth structure (potentially causing mismatched points via occlusion). It is not clear to what extent natural stimulus variability (for example, in the number of mismatched points, or available reference planes) impacts binocular correspondence and therefore depth perception in real scenes.

There is some basis for evaluating the available information. Evidence suggests that the spatial frequencies most important for stereopsis are middling spatial frequencies between 2 and 8 cycles per degree (Frisby and Mayhew 1978, Badcock and Schor 1985, Lee and Rogers 1997). However, in real scenes large surfaces low (even lacking) in contrast can usually be grouped to some distal feature. Measures of disparity acuity for isolated targets will necessarily be local measures. Conversely, distance acuity in real scenes will almost certainly depend on more global information content. Therefore, the relevant 'local contrasts' are not well defined. It is possible local contrast is a relevant parameter for both binocular and monocular acuity; but in the context of natural scenes it is an empirical question. In real scenes, perceptual grouping mechanisms allow for inferences about surfaces and objects that are embedded in a broader context. That is, nonlocal image features are informative about the task. With abstract (cue reduced) stimuli, these sorts of contextual distance inferences are not always possible.

Even so, disparity discrimination thresholds provide a relatively believable laboratory estimate of the reliability of disparity information. For the purposes of modeling disparity acuity I have adopted Colin Blakemore's (1970) exhaustive measurements of the space. These measurements at least capture the substantial impact of peripheral estimation. Estimating the overall reliability of monocular information is in

some ways much simpler because of its immediate relationship to retinal images, and in some ways much more difficult because its stronger dependency on poorly understood prior assumptions.

Monocular Acuity – Relevant Experimental Findings:

Greenwald and Knill (2009) measured the relative reliability of monocular and binocular cues for slant estimation as a function of eccentricity and position with respect to the horopter. However, these measurements were made in near space (grasping range). Further, slant and distance discrimination are certainly related, but it is not entirely clear how to predict distance discrimination thresholds from slant discrimination tasks. Finally, the textures and bounding contours of the surfaces were parametrically varied from trial to trial, but they were drawn from fixed statistical class of qualitatively identical texture and contour.

Maloney and Landy point out (1989) that the reliability of texture cues will depend highly on the particulars of the texture. There has been some effort to quantify in what way. For example, it is known that the foreshortening of isotropic texture elements is a stronger cue than the change in density of texture elements with slant (Knill 1998). This was consistent with the results of a plausible ideal observer. However, for textures with substantial relief (e.g. cobblestones), there will be less foreshortening, and thus density will be the only available information (Saunders 2003). Therefore, it is conceivable that the relative importance of monocular cues could be sensibly modeled if the likelihood of various textures and surface reliefs were known.

There has been almost no attempt to quantify the monocular information available in natural scenes. Yang and Purves (2003) did measure the relief of the ground plane using a scanning laser range finder, although they did not relate it closely to textural

information (Yang and Purves 2003). Their dataset is one of the only other data sets existing the public sphere that relates measurements of real scenes to their images, and it was collected with much older technology than the dataset collected here. Without direct measurement of the relationship between images of scenes and the structure of scenes no model can predict the average influence monocular depth cues. Monocular cues other than textural cues have been studied extensively in the lab; however extrapolating from these laboratory studies to natural scenes is constrained by a lack of understanding of how cues of various reliabilities are distributed in real scenes. The frequency with which various monocular cues of varying reliability are available in natural scenes is not known. It is not even known how to identify the boundaries of the relevant surfaces or textures in the model outside the context of an experimenter's artifice.

1.4 - My Approach

As an alternate approach, I have decided to measure distance discrimination thresholds across a large sample of natural scenes by simulating them in the lab. This approach has the advantage of making direct measurements of discrimination performance with the real scenes, while still conferring many of the logistical advantages of running experiments in the lab. To this end I have designed a unique virtual reality display built to maximize the similarity between the visual stimulus and the measured real world scene. Thereby, psychometric functions can be measured reflecting average performance across a large sample of real scenes.

Mechanistic models that make stimulus-general predictions for performance are clearly a worthwhile goal but they are not always tractable to develop. Here I emphasize the value of descriptive measurements of performance in natural scenes as a first approximation for baseline acuity. Measurements of this sort can themselves be

predictive both by being repeatable measures across similar experiments as well as by parameterizing models meant to test other mechanistic predictions (e.g. the cue combination rule).

Virtual Reality:

Running psychophysical experiments using computerized displays has considerable logistical advantages over running psychophysical experiments in real scenes, e.g. outdoors. One major advantage is the potential for running a large number of trials with diverse stimuli. Another advantage is the availability of (and experimental control over) an exact generative model for the stimulus. Knowledge of the generative model allows for precise a specification of the psychophysical task; thereby improving the ease with which ground-truth accuracy can be related to the available sensory information. Philosophically, I would rather run experiments in real scenes, however the logistical disadvantages are too difficult to overcome.

A displayed stimulus reaching sufficient consistency with the depth structure of its associated physical scene can be considered a 'virtual' reality (VR). At some limit of consonance between the displayed stimulus and the image of a real scene, virtual reality would presumably be indistinguishable from actual reality. Thus virtual reality experiments can plausibly serve as a surrogate for experiments run in the physical scene being simulated. In common usage, virtual reality has been associated with head-mounted displays (HMDs), however the term applies more generally to any attempt to replicate the sensory experience of reality.

For my depth acuity benchmark, I propose a VR stimulus that has a higher degree of cue consistency with real scenes than previously documented in the psychophysical literature. I have some caveats to this claim. HMDs have the notable advantage of

accounting for parallax. Volumetric displays (Love, Hoffman et al. 2009) have the advantage of accounting for the large defocuses that occur at near distances. However, I believe the study described here to be the most rigorous attempt at replicating real images of real physical scenes for which the ground-truth depth structure is known.

In so far as it is not possible to technically replicate some relevant aspect of the image, rigorous attempts should be made to design experiments that limit the relevance of the technical shortcoming (see Chapter 3). That is, the explicit goal of this approach is to perfectly replicate a realistic image to the extent technically possible. Therefore, I can run an experiment in the lab, and have a reasonable *naturalistic proxy* for an experiment run in the natural world. I achieve this by directly measuring real stimuli, and displaying them in a fashion consistent with a stationary (head fixed) view through a window (the stereoscopic display).

While these conditions are a more restrictive than natural viewing, I believe there are major advantages gained. In particular, I believe that these conditions are less drastic, and more naturalistic, than severe restrictions on the stimulus set. It is known that there is rich information available in textures, complex bounding contours, and general prior knowledge about images. Typical virtual reality displays use computer graphics to render their scenes; and graphics models are limited in their scope. It is not known what specific image properties impact their realism; consequently these synthetic images rarely appear 'real' by anyone's definition. Certainly not as real as those images in Figures 4 and 5.

Figure 4 – A Stereoscopic Image from my Database

A sample stereoscopic image from my database. Even at print resolution of 300 dots-per-inch these were too large to each fit on their own page. They were scaled down ~4-1.

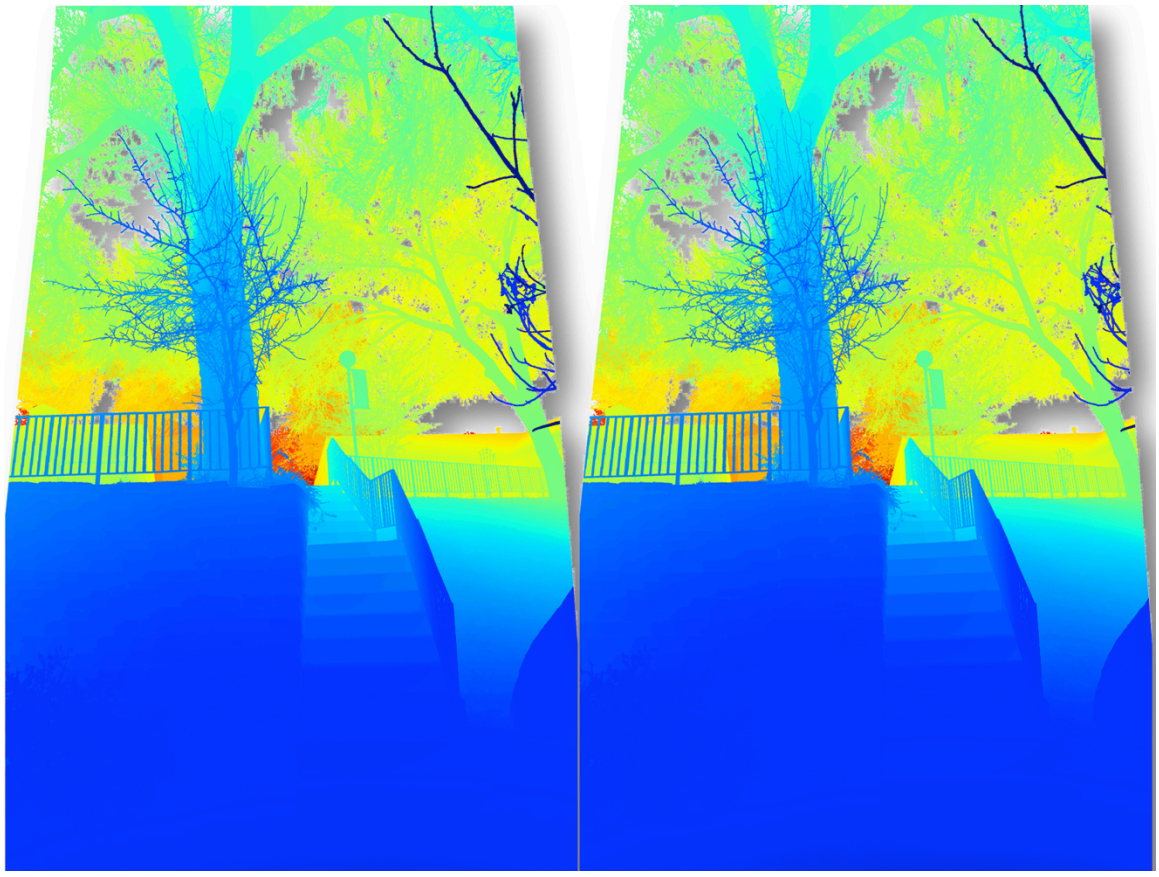

Figure 5 – Corresponding Range Images

These are range maps that correspond to the stereoscopic images in Figure 4. Cooler colors denote nearer objects. Notice that the two range images correspond to the two viewing locations, accounting for changes in the occlusion pattern. Transparency (shown by shadows) denotes non-responses from the rangefinder.

It is thought that humans may have detailed knowledge about scene structure not captured by the graphics model. Thus, observer performance may be impacted by deviations from realism. Accepting as a limitation head fixed viewing of static scenes through a windowed aperture allows me to make essentially perfect reproductions of real

scenes. That is, the advantage I gain is freedom from relying on computer graphics for my generative process. The real world generates my stimuli for me.

Defocus:

While my goal is to perfectly replicate the stimulus produced by real world scenes, this is not technically possible. For one, the eye can adaptively change the power of its lens to accommodate to different distances. Vergence-Accommodation conflicts have been shown to create discomfort in stereoscopic displays(Shibata, Kim et al. 2011). Further, blur has been shown to be a potential complementary cue to disparity(Held, Cooper et al. 2012). However, the mathematics of defocus is closely related to the mathematics of disparity. In particular, objects a large distances will not vary much in their defocus, whereas nearby objects will vary more dramatically. Most of these experiments are run with much closer base distances than under investigation in my study. Objects in my collected scenes were never closer than 3m. Similarly, the window was placed at this (relatively large) distance. Therefore, the largest defocus in the image will be on the order of 1/3 diopter, not much larger than defocus detection thresholds measured for patches of natural scenes (Sebastian et al. in press).

Parallax:

Parallax is an important cue, however it is intractable to measure scenes and images from all possible viewpoints. Fixing ego-motion allows for a perfect replication of the stimulus, rather than a reconstruction from multiple viewpoints. While head fixed viewing is a bit unnatural (akin to controlling fixation) the tighter relationship with the real scene is worth the trade off. Since the aperture is a window straight ahead along

primary gaze, uncomfortable gaze angles are not necessary. Therefore, this restriction of the head is somewhat reasonable.

Signal Detection Theory:

One of our goals as psychophysicists is to characterize the fidelity sensory estimates. Usually, estimation precision is more meaningful than estimation accuracy. This is because estimation biases can often be corrected by the motor system, whereas imprecise estimates cannot. Consequently, it is common in perception research to focus on acuity measures, which reflect the limit of sensory precision.

The fidelity (and bias) of these estimates is often measured by obtaining psychometric functions. A psychometric function for depth discrimination is illustrated in Figure 6a. The horizontal axis gives the depth of a comparison stimulus assuming a standard stimulus at a depth of 10 m. The vertical axis givens the proportion of times the observer reported the comparison was judged to be nearer than the standard.

The precision of an estimate is typically defined to be the standard deviation of a cumulative Gaussian fit to the psychometric data. This definition is based on signal detection theory (Figure 6b,c). See *Signal Detection Theory and Psychophysics* (Green and Swets 1966) for review. According to signal detection theory, each stimulus level elicits a distribution of neural activity (solid and dashed curves in Fig. 6b,c). The signal-to-noise ratio, d', is the number of standard deviations between these two distributions. Thus, if threshold (a just noticeable difference) is defined as the difference between the standard and comparison stimuli where the signal-to-noise ratio is 1.0 ($d' = 1.0$), then that difference is the standard deviation of the Gaussian fit to the psychometric function.

(For the yellow and green points in Figure 6a the corresponding two distributions in Figure 6c are separated by one standard deviation.)

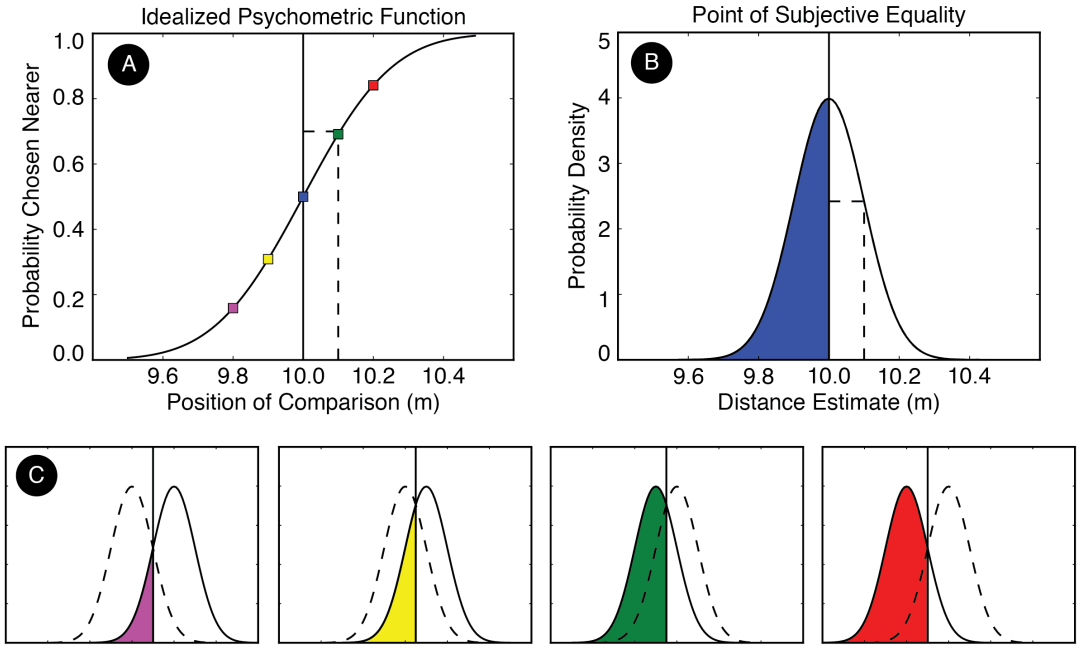

Figure 6 - Schematic of the Standard Signal Detection Theory Model

(A) An idealized psychometric function with a plausible threshold for the distance discrimination task at an absolute distance of 10m. The solid vertical line shows the point of subjective equality, while the dashed line shows the ~70% threshold. **(B)** The underlying signal detection theory model for decisions made at the point of subjective equality. The solid line shows an unbiased criterion, while the dashed line shows the standard deviation of the Gaussian estimate. Note the equivalency with threshold. **(C)** The underlying signal detection theory model for easier discriminations (d'=1,2). The dashed curve reflects the distribution of distance estimates for a standard fixed at 10m. The solid curve reflects the distribution of estimates for varying comparison distances.

The shaded areas correspond to the probability mass associated with the data points in 'A'.

Formally Modeling Depth Perception:

In order to describe my psychophysical paradigm it is useful to have a more formal model to address. Here, I describe the standard cue combination paradigm, so that I can address how my experiments relate to the literature. I cast my model specifically with regard to the problem of distance estimation, but a similar framework can be (has been) applied to similar tasks, e.g. estimating surface orientations. My particular emphasis is on how to generalize the cue combination framework to more naturalistic circumstances.

The psychophysical task is to determine the nearer of two locations in the scene. Therefore the task can be conceived as determine the sign of the difference of two distance estimates. We assume distance estimates are distributed as a Gaussian, therefore their difference will also be a Gaussian distributed estimate. Since we assume distance estimates are unbiased, this depth estimate will be centered on the correct depth, and the criterion will be placed at zero. That is, the only influence over the position of the decision variable's distribution is the actual difference in distance denoted Δ_d. Thus, the mean of this distribution is known objectively under the model since the physical distances present in the scene are known objectively.

Of particular interest is the standard deviation of these depth estimates. In the signal detection framework, the standard deviation of the depth estimates can be measured via threshold denoted by σ. It is known that sigma can be influenced by many other difficulty parameters including viewing conditions, observer identity, and image

content. In other words, threshold can be thought of as being a function of a high-dimensional parameter vector, i.e. $\sigma\left(\vec{P}\right)$. Thus, the resulting decision variable $\hat{\Delta}_d \sim N\left(\Delta_d, \sigma\left(\vec{P}\right)\right)$.

Since we are interested in a measure of acuity to use as a baseline measurement across observers and images, I prioritized two parameters of difficulty tied to the viewing configuration, specifically the absolute distance at which the distances were discriminated, denoted d, and the visual angle separating the targets, denoted θ. Thus, we are only considering threshold to be a function of these two parameters, i.e. $\sigma(d,\theta)$.

Clearly there are other interesting parameters to vary in the difficulty parameter vector, notably properties of the image vector $\vec{I} \subseteq \vec{P}$. It is known that image properties could impact the difficulty of the task (see above). Rather than hold these parameters fixed on every trial, we opt instead to marginalize across the difficulty parameters to the extent possible. For further discussion of this approach, see the next section 'Naturalistic Psychophysics.'

One common approach to understanding mechanisms of depth perception is to introduce a set of cues represented as functions over the parameters. Formally, a cue can be thought of as a functional relationship $\hat{\Delta}_c = f_c\left(\vec{I};\vec{P}\right)$. For some cues, particularly disparity, sensible definitions for f_C (algorithms for getting distance estimates from images) have been proposed cues (Marr and Poggio 1979). This is a useful approach; f_C will therefore have a defined value for any given input.

A considerable advantage to using cues is that the dimensionality of the problem is reduced. Marginal likelihood distributions of the form $p\left(\hat{\Delta}_c|\Delta_d\right)$ are much easier to estimate than the full joint likelihood distribution, i.e. $p\left(\vec{I},\vec{P}|\Delta_d\right)$.

Another advantage of this approach is that it is easy to design cue-isolating stimuli. The responses of multiple such functions f_C can be made independent of each other over the experimental stimulus set.

Since it is known that multiple cues are used to estimate depth, it is usually presumed that there is some 'rule of combination' by which the multiple estimates $\hat{\Delta}_c$ can be combined. The 'Psychophysical Observer' (Maloney and Landy 1989) is designed to provide a psychophysical test for the combination rule. In it, stimuli are designed such that the cues (or algorithms like f_C) respond independently to different isolated components of the stimulus. Thus, individual psychophysical experiments can be run to determine the variability associated with $p\left(\hat{\Delta}_c|\Delta_d\right)$, (the acuity for the cue), thereby creating independent measures of the parameter in their combination model. As a consequence of the independence inherent to the stimulus design, a prediction for the combined stimulus can be directly tested.

Here we model two cues, monocular and binocular cues to depth denoted with the subscripts m and b respectively. Assume that the independent depth estimates generated by the cues are normally distributed, unbiased, independent distance estimates, i.e. $p\left(\hat{\Delta}_m|\Delta_d\right) \sim N\left(\Delta_d, \sigma_m(d,\theta)\right)$. Under this set of assumptions, the optimal combination rule is a linear weighting of the independent estimates. Formally:

$$\hat{\Delta}_d = \frac{\sigma_m^{-2}(d,\theta)\hat{\Delta}_m + \sigma_b^{-2}(d,\theta)\hat{\Delta}_b}{\sigma_m^{-2}(d,\theta) + \sigma_b^{-2}(d,\theta)}.$$

Under this particular assumption set, the final estimate will still be normally distributed, i.e. $\hat{\Delta}_d \sim N\left(\Delta_d, \sigma(d,\theta)\right)$. The estimates variability is equal to $\sigma^2(d,\theta) = \frac{1}{\sigma_m^{-2}(d,\theta) + \sigma_b^{-2}(d,\theta)}$. Here I use Blakemore's (1970) data to as an estimate of $\sigma_b(d,\theta)$, while I directly measure $\sigma_m(d,\theta)$ and $\sigma(d,\theta)$ in order to test the model.

Naturalistic Psychophysics:

There is continued debate about the advantages and disadvantages of using natural images as stimuli for studying vision. Recently, the debate is over best practices for physiological recordings of neurons in visual cortex (Rust & Movshon 2005, Felsen & Dan 2005), however this is an old debate in behavioral sensory psychology. There are merits to all the standard arguments. Ultimately, the appropriateness of naturalistic experiment design depends on the scientific question being asked.

It is frustrating that the term 'natural' is rarely operationalized. Many factors intuitively contribute to the naturalism of an experiment. Important factors include the choice of task, images, participants, viewing conditions, or reporting methods. Like most arguments of this sort, extreme positions are untenable. Even the most natural designs are still experiments. Even strict experimental controls are tempered by natural variability. We would like to predict all of the data, irrespective of naturalism. The value in operationalizing what we mean is that we can talk about the scientific purposes served by the specific design choices we made.

All experiments involve choices about the design. Some parameters need to be controlled. Certainly, there needs to be control over (i.e. experimenter knowledge and explicit sampling of) the parameters whose influences are directly under study. Other parameters may or may not be controlled. They potentially influence the value of the dependent variable, but their influence is thought to be independent. Thus, there is no principled reason to pick any particular level of the unspecified parameter with respect to the model. There is just an arbitrary choice that needs to be made.

In visual psychophysics a frequent example is the choice of visual stimulus, i.e. the particular pattern of light and dark on the screen. Usually the model being tested is of

substantially lower dimensionality than the stimulus. Thus, there are many stimulus dimensions unconstrained by the demands of the experiment. Traditionally, these additional stimulus dimensions are either held fixed (e.g. Gabor patches), or are sampled randomly according to some known statistical distribution (e.g. noise samples). In the experiments presented here the stimuli from trial to trial are randomly sampled with minimal conditioning from my database of natural scenes. This choice was made because the explicit goal of the study was to provide an estimate of overall sensory acuity. There is a simple argument for allowing trial-to-trial variability in this context. Stated succinctly, averaging is good.

It has been suggested to me that this sort of an approach adds noise to my measurements because, 'threshold is changing from trial to trial.' It is not entirely clear what is meant by threshold in this context. There is no way to estimate a threshold from a single trial. Presumably what is meant is that parameters known to impact performance in other depth discrimination experiments are left to vary as they do in natural scenes. That is undoubtedly true, but arguably correct practice in an acuity experiment where the only mechanisms of interest are those that measurably influence average performance.

It has further been suggested that sampling randomly from natural scenes is improper because natural scenes constitute an unknown statistical distribution. Therefore, it is impossible to know if the particular sample used in the experiment is biased. While this may be true, within the context of the experiment any and all parameters of the sampled stimuli are well defined. It is true that a second, similar experiment will use a different sample from natural scenes, and therefore may produce slightly different results. However, the problem is worse when arbitrary choices are made once and then held fixed.

I ran a simple simulation to illustrate my argument. Suppose, as my detractors may assert, that the standard deviation of the decision variable is random on a trial-to-trial basis. I decided to use the Bayesian conjugate prior for variances, i.e. a distribution from the scaled inverse gamma family. Rather than making any deep formal argument, suffice it to say this is a probability distribution that does what you might expect. It does not go below zero, but is skewed highly towards larger variances. It is parameterized by two values, one that determines the overall tendency of threshold (the acuity measure we are interested in) and one that determines how variable 'threshold' is from trial-to-trial.

Those opposed to naturalistic designs suggest that it is important to hold properties of the image fixed from trial-to-trial in order to measure a stable threshold. I operationalize this as taking a single sample from this distribution of thresholds, and measuring an entire psychometric threshold with the difficulty fixed at this arbitrary level. Not surprisingly, the measured threshold is entirely determined by this single sample taking from the threshold distribution.

Alternatively, I propose allowing difficulty to vary on every trial as it would in natural scenes. Therefore, the standard deviation parameter associated with the decision variable in the simulation varies on a trial-to-trial basis. The result is that the threshold measured in the simulated experiment is largely determined by the parameter in the distribution of thresholds we are interested in. That is, a measure of central tendency of the threshold distribution. Better yet, the variability across repeated experiments (a simulated meta-analysis) shows that the value measured in the naturalistic experiment is not appreciably more variable from experiment to experiment as a series of experiments in which threshold is held perfectly fixed. Whereas, if stimulus properties (and therefore threshold) were held fixed from trial to trial, but varied from experiment to experiment,

the simulated meta-analysis reveals that this approach yields much less stable measures, as would be expected from a sample, rather than sampling distribution. It may be a simple point about the advantages of using natural stimulus variability for acuity measurements, but one worth making nonetheless.

Summary:

With the current work I take a different approach, combining the naturalistic stimuli of in-field studies with the rigor of laboratory studies via virtual reality. The current work accomplishes three overall aims: (i) to build a high quality natural scenes dataset pairing stereoscopic images with range images, (ii) to bring these images into the lab in order to measure naturalistic depth acuity, and (iii) to demonstrate (with disparity) how these measurements can be used as a benchmark in order to evaluate the relative importance of various cues to depth.

The first aim is inherently useful. Understanding depth perception ultimately entails understanding the relationship between sensory stimuli (i.e. stereoscopic images) and the physical scene (i.e. 3D structure). Thus, this database is useful not only for my subsequent psychophysical aims, but also as direct measurements useful for developing a generative model of images. The creation of this database is a necessary next step in advancing theories of depth perception. This aim involved substantial technical efforts. Methodological details are outlined in the next chapter (Chapter 2).

The second aim answers a fundamental question about depth perception that has not yet been answered in the literature. What is average human acuity for depths in real scenes? How does it vary with viewing geometry? While disparity discrimination thresholds have been measured, discrimination thresholds in real scenes containing rich pictorial cues to depth have not. Estimation designs have been run outdoors, but they are

not as sensitive as discrimination designs for measuring the precision of sensory estimates. Ultimately, the precision of sensory estimates is what limits an observer's capacity to learn accurate sensory-motor mappings. Thus, the second aim provides a descriptive measure the performance of people in a typical task requiring a depth estimate. Knowing the precision of depth estimates in real scenes provides a benchmark against which the performance of models or other observers can be compared. Methodological details for the second and third aims can be found in chapter three.

The third aim is more exploratory in nature. Most laboratory depth perception experiments go to great lengths in order to isolate the cue of interest. Thus, these experiments convincingly demonstrate that the isolated cue is sufficient to evoke a depth percept. However, it is not clear from these experiments that the absence of the measured cue would substantially impact depth perception when a different image associated with different cue reliabilities is used. For example, Palmisano et al. (2010) shows convincingly that binocular disparity is a viable cue to depth even at quite large physical distances. Disparity was the only cue to depth available, and observers could discriminate depths. However, it does not show that disparity is a substantial contributor to depth perception at large distances when other more reliable cues are available to substitute. Using natural images as my stimuli allows me to measure the average usefulness of disparity as a cue to depth as a function of distance and the visual angle separating the targets. Since I am beginning with natural images, and it is unknown what cues they contain I find that the approach is more analogous to removing an information source than isolating one. However, under the cue combination model, the math is equivalent.

Here, I focus in particular on monocular performance (i.e. performance without binocular disparity). In principle this approach could be applied more generally. The utility of this approach is twofold. First, it is a direct measurement of what performance is likely to be in the presence of a visual deficit (e.g. the loss of an eye). Second, it provides an estimate of the reliability of the remaining cues that could be used in conjunction with the standard cue combination assumptions to estimate the acuity loss accounted for by the missing information.

Chapter 2: Construction of a Database

2.1 – Collection Overview

Image collection involved integrating sophisticated equipment, and preparing it for use in the field. Distance was measured using a Riegl VZ-400 scanning laser range finder chosen for its state-of-art capture density and precision. Images were collected using a Nikon D700 digital SLR camera. Both of these devices were mounted on a custom built robotic gantry to control positioning (see Figure 7). All of the equipment was controlled by a laptop running custom software written in-house. Each scene consists of at least two camera images spaced 65mm apart (average distance between the two eyes for males) and a corresponding stereo pair of range scans.

Outdoor image collection introduced significant constraints. The entire apparatus had to be battery operated. The equipment was extremely heavy, and choice of wheels proved critical to successfully navigating terrain. One limitation of the apparatus is that images needed to be captured sequentially. Outdoors, conditions such as wind, lighting, and animal life were not under immediate control. Since the collection process took significant time for each scene (on the order of minutes), object motion and lighting changes had to be avoided by waiting, persistence and post-hoc selection. Therefore, moving objects (e.g. cars, animals, wind, or pedestrians) were selectively avoided as much as possible.

One major advantage of our dataset is that the positioning robot allowed for precise colocation of the camera and the range finder by repositioning the apparatus. Without repositioning the quality of the dataset would have been reduced by what are known as “half-occlusions,” resulting from the separation between the nodal points of the range finder imaging system and the camera’s lens. Because different viewing locations

result in different patterns of occlusion, there would be portions of the range images not visible from the camera's viewpoint, while there will be portions of the camera image not visible from range finder's viewpoint. The range scans and camera images were captured from as close the same location as possible, thereby minimizing half-occlusions in our scenes.

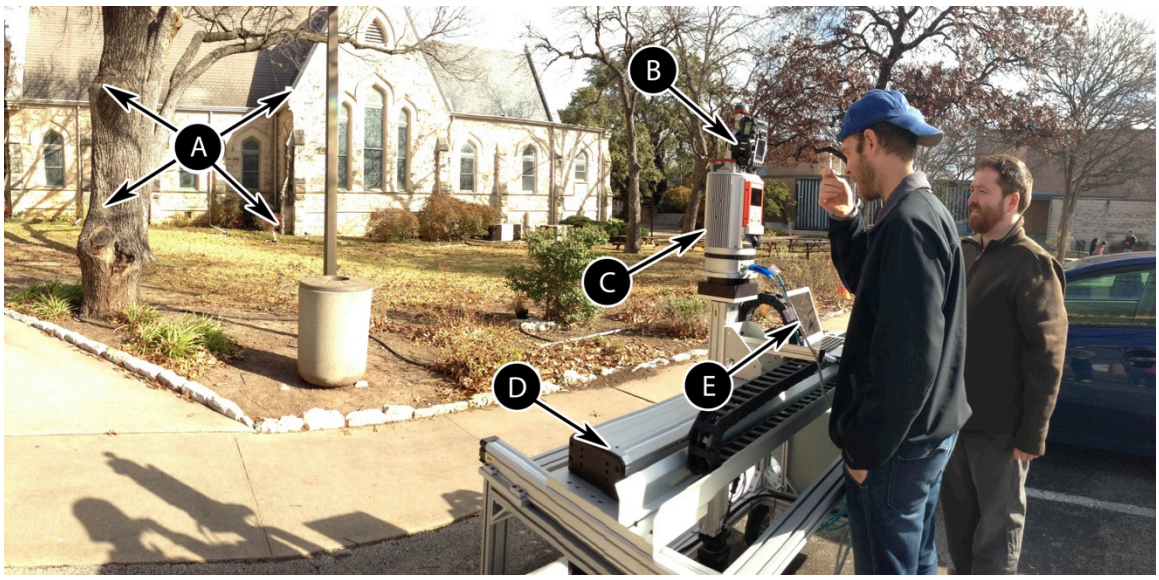

Figure 7 – Natural Scene Measurement Apparatus

(A) The scene being measured. **(B)** The Nikon D700 Digital SLR Camera. **(C)** The Riegl VZ-400 Scanning Laser Range Finder. **(D)** The custom built positioning robot. **(E)** Laptop running custom control software.

Image content was mixed. In order to stay consistent with natural viewing conditions, the images were captured from roughly eye height. As mentioned, viewing locations were heavily constrained by the bulkiness of the equipment. All images were

taken within walking distance of the Psychology building on the University of Texas at Austin campus. Images included buildings, signs and fences as might be expected. An effort was made to include natural features like trees and foliage. Some materials are not conducive to range scanning because they either absorb too much infrared (e.g. tires), or reflect it away from the camera in a specular manner (e.g. glass). To the extent possible objects of this sort were avoided as the subject of photographs. A modest effort was made to capture objects at multiple depth planes, with no objects closer to the observer than 3m. While this introduces the possibility of some global photographer's bias, the points in the images selected for depth discrimination were randomly distributed throughout the images.

The vertical field of view of the Riegl VZ-400 is fixed at 100° because of the rotating mirror scanning mechanism. The horizontal field of view is arbitrary, however 60° was chosen because that closely matches the camera's horizontal field of view. The density of the scanned point cloud can be changed as a trade off with scanning time. A density of .04° per sample was chosen as a compromise between scanning speed and resolution. This is approximately half the linear resolution of the image, in which a pixel subtends roughly .02° of arc.

It is worth noting that the scanner's samples are relatively evenly spaced as a conic projection of the scene, rather than the standard planar projection used in the camera. That is, azimuthal angles are evenly sampled (unlike with planar projections), and field of view is biased towards the upper visual field. Therefore, there is not a perfect overlap between the projections. Further, the vertical scan lines produced by the range finder are sampled with arbitrary phase, and thus are not trivially mapped to the

camera image, nor any rectilinear grid. An extensive review of the geometric registration procedure I developed can be found in the following section (2.2).

The Nikon D700 camera body mounted on the scanner had a 20mm fixed focus lens. Changing the focus changes the projective geometry, so the fixed focus lens was used to keep the geometry as consistent as possible on an image-to-image basis. However, this fixed focus lens should have little impact on the focus quality of the images as no objects were photographed from a distance of less than 3m and the camera's aperture was always very small (~2.5mm diameter at a fixed F-Number of 8.0).

Images were captured in 14-bit "raw" Nikon Extended Format (NEF) allowing for a large dynamic range and responses linear with veridical luminance. The spectral response properties and linearity of the camera have been characterized with the associated methods and details presented in section 2.3. The camera's gain was always set to the standard International Standards Organization (ISO) 200, with apertures as small as possible to achieve reasonable exposure over maximum open-shutter durations of 10ms. Care was taken to minimize the complementary metal-oxide semiconductor (CMOS) sensor response clipping. The pixel dimensions of the images measure 2844 x 4284 pixels.

Note that the images were captured in portrait orientation with a 10° backwards slant. Therefore, the camera's image plane was aligned with the rangefinder's field of view. The imaging geometry is a bit involved. The next section (2.2) goes into more detail about the geometry intrinsic to the camera. Section 2.4 describes correcting the intrinsic geometry appropriate for display and analysis.

2.2 - SPATIAL CALIBRATION

A spatial calibration of the range finder was performed in the factory. Unfortunately that calibration was measured for compressed format images in order to work with their graphical user interface. Therefore, in order to preserve the 14-bit depth of the raw images, it was necessary to develop a custom calibration, which communicated with the device through the application programmer interface.

2.2.1 - The Camera Model

The calibration I performed was designed to mimic the calibration described by Riegl. The linear (standard pinhole model) component of the full camera model can be made formal as follows:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \begin{bmatrix} \alpha_x & 0 & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{xx} & r_{yx} & r_{zx} & t_x \\ r_{xy} & r_{yy} & r_{zy} & t_y \\ r_{xz} & r_{yz} & r_{zz} & t_z \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}$$

where the column vectors reflect the point in camera and world coordinates respectively, the 3x4 matrix reflects a rotation and translation between the coordinate systems the estimation of which is described in the next chapter, and finally the 3x3 matrix is referred to as the intrinsic camera matrix. In the intrinsic camera matrix (α_x, α_y) describes a scaling appropriate for the focal length of the camera, while (u_0, v_0) is referred to as the principle point and reflects the intersection of the optic axis with the image plane.

Additionally, we model higher order distortions caused by the camera's lens system. Specifically, we model four radial distortion coefficients, and two tangential distortion coefficients. We begin by defining some convenience variables (u_d, v_d) which result from a perspective pinhole projection of the point in the camera's coordinate system. That is, $u_d \triangleq \frac{x_c}{z_c}$, and $v_d \triangleq \frac{y_c}{z_c}$. To simplify the resulting expression, we define further convenience variables (x', y', ρ^2) by: $x' \triangleq \frac{u_d - u_0}{\alpha_x}$, $y' \triangleq \frac{v_d - v_0}{\alpha_y}$, and $\rho^2 \triangleq x'^2 +$

y'^2. Finally, we can apply the lens distortions and describe the pixel coordinate, (u, v), of the projected 3d point such that:

$$u = u_d + x'\alpha_x(k_1\rho^2 + k_2\rho^4 + k_3\rho^6 + k_4\rho^8) + 2\alpha_x x'y'p_1 + p_2\alpha_x(\rho^2 + 2x'^2),$$

$$v = v_d + y'\alpha_y(k_1\rho^2 + k_2\rho^4 + k_3\rho^6 + k_4\rho^8) + 2\alpha_y x'y'p_2 + p_1\alpha_y(\rho^2 + 2y'^2),$$

where the k parameters reflect the four radial lens distortions and the p parameters reflect the two tangential lens distortion parameters. These equations together describe the entire camera model.

The parameters of this model were estimated in two stages. The first stage estimates the value of the parameters determined by the imaging geometry of the camera itself, independent of its physical orientation in the scene. These include the intrinsic camera matrix and the nonlinear lens distortion parameters. The rotation and translation which best aligns the data was estimated separately on a per-image basis.

2.2.2 - Estimating the Intrinsic Camera Parameters

The parameters intrinsic to the camera's imaging geometry were estimated on their own, and then fixed across all the images. These parameters were estimated after exploring a number of techniques. The strategy ultimately chosen was one most closely resembling the method described by Zhang (2000). Zhang's method relies on taking images of a slanted planar grid from multiple viewpoints to provide sufficient algebraic constraint on both the intrinsic camera matrix, and the physical orientation of the camera in each image(Zhang 2000). One small difference between the method of Zhang and the one used here is that the lens distortion correction described in Zhang's paper is less extensive (uses fewer distortion coefficients) than the model described above.

The large field of view, and fixed focus of the lens made it challenging to create a suitable grid of targets. A poster format printer was tried. This approach yielded a grid

with thousands of easy to detect features, however even this was relatively small, and therefore needed to be about one meter away to cover the camera's field of view. At such a small distance, defocus is introduced in the image. Additionally, the paper was difficult to perfectly flatten out, adding noise to the method's assumption of planarity. I also tried using the tile floor near our offices. These provided a large enough stimulus to impose fewer constraints on the viewpoint, and it was at least as convincingly planar. However the edges were lower in contrast, and the lighting was more difficult to control. The irregularity of the features made automated feature detection impractical. I opted to hand localize the corners of the tiles. Since the feature locations were hand selected, there was probably some (difficult to measure) human error introduced by this method. However, Zhang found that the larger slants made practical by this method provide more robust estimates of the orientation of the plane. It was difficult to assess whether error from hand localization was larger than the error from introduced by holding the poster roughly parallel to the image plane (necessary to fill the field of view). Ultimately the data from both approaches were combined to create the best estimates of the camera geometry.

Zhang's method uses sophisticated geometric abstractions beyond the scope of this document. Suffice to say that the maximum likelihood parameter estimates were found by minimizing the L2 norm of the residual. All model parameters were optimized simultaneously using the Levenberg-Marquardt nonlinear minimization algorithm. Levenberg-Marquardt minimization requires an initial guess. Zhang's paper goes into detail on how this initial guess can be obtained for the intrinsic camera matrix. For the lens distortion parameters, an initial guess of zero was used. Calibration results were excellent, suggesting an average calibration error of around 1-2 pixels. That said, during

this entire calibration procedure it is difficult to have complete confidence in ground truth estimates of error, since many sources of noise challenge the assumptions in the model.

2.2.3 - Estimating the Rotation and Translation Parameters

The rotation and translation parameters were allowed to vary on a per-image basis. In theory, the robotic positioning and precision mounting of the camera should not require an image-by-image registration. However in practice it noticeably improved performance in the image registration by at least a couple of pixels on average. Note that each pixel corresponds to roughly one minute of arc, and therefore a couple of pixels of rotation seems a like a reasonable error to expect attributable to vibrations in the rig during field work.

For the hand registration a piece of software was developed to display raw images and range data side by side. The software allows for the user to zoom in on individual pixels to improve the reliability of the clicks. I clicked on at least 20 pairs of matched key points for each image. I tried to choose points that appeared stable, and are obvious in both the range and camera images i.e. rigid objects unlikely to be affected by wind. For some images finding reliably matching points of this sort was difficult, or objects were clustered biasing spatial sampling in the image, adding substantial uncertainty to accuracy measurements. In all, registration appeared generally quite good. Close visual inspection suggested that the final registered images again had errors on the order of a pixel or two, with errors being largest towards the edges of the image which were not included in the cropped images used for subsequent projects.

2.3 - LUMINANCE CALIBRATION

For the purposes of this database it was important that we not only have a precise spatial calibration, but also precise luminance measurements. To be more precise, we wanted an accurate linear mapping from red, green, and blue (RGB) camera responses into standard color spaces the XYZ color space developed by the international commission on illumination (abbreviated CIE from the name in French). This was accomplished by measuring the camera's responses to patches of monochromatic light covering the visible range, and then fitting a transformation between these responses and known color spaces.

2.3.1 - Measuring Monochromatic Camera Responses

The first step in performing the luminance calibration was to measure the responses of the camera to monochromatic light. An optical bench was used to position the camera, a monochromatic illuminant, a white reflectance standard, and a spectroradiometer in an otherwise dark room. A schematic of their physical arrangement can be seen in Figure 8.

The monochromatic light source illuminated a Labsphere certified flat white reflectance standard with a narrow pass band of light. The average wavelength was stepped in 5nm increments over the entire visible range from 400nm-700nm. Note that the camera and spectroradiometer are at an equal angle relative to the reflectance patch and the illuminant. The reflectance standard is highly Lambertian. Therefore the luminous intensity will be the same as observed by the two devices according to Lambert's cosine law (Lambert 1760). The spectroradiometer provides a measurement of the spectral irradiance of the white test plate for a given wavelength setting on the monochromatic source.

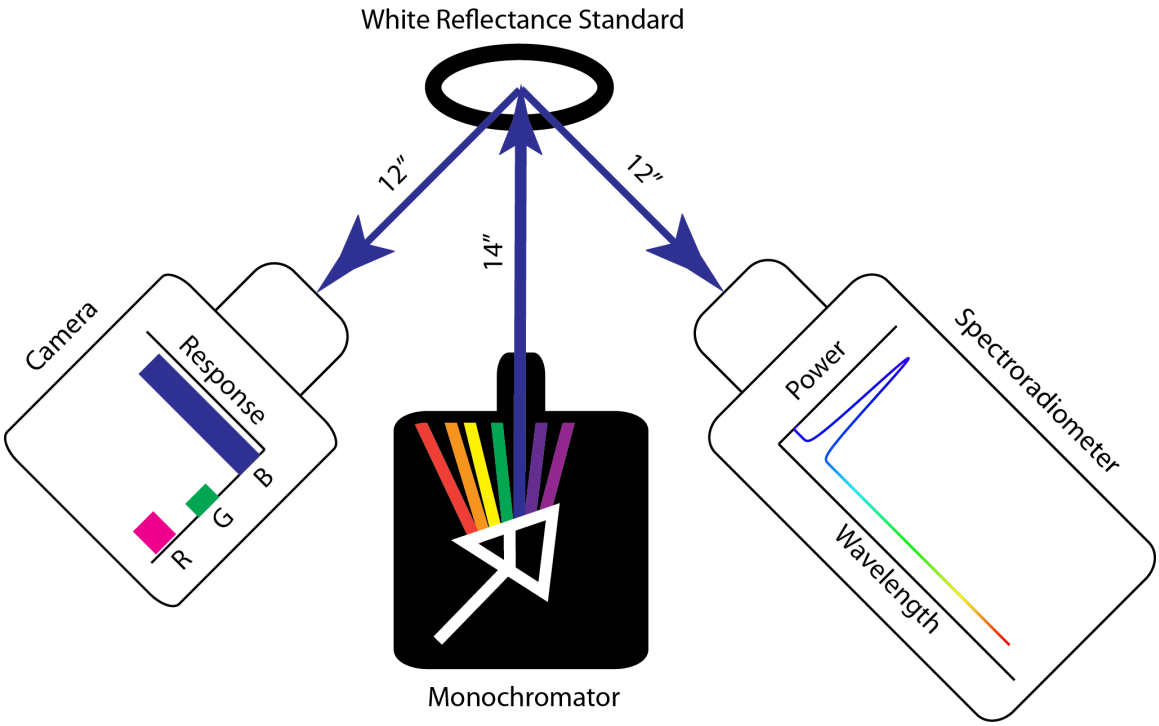

Figure 8 – Schematic Overview of Color Calibration

The D700 camera and a spectroradiometer measure responses to monochromatic light. These camera sensitivity measurements can be used to map between camera responses and color spaces from the literature.

The integrated power will depend on the area of collection and collection time. The camera's aperture was widened to its widest setting, setting the f-stop to f/1.8. Shutter speeds were allowed to vary across wavelength in order to produce robust CMOS responses. Open times were recorded with the metadata for each image for post-hoc

analysis in the extended information file (EXIF file). Responses will also depend on the ISO setting that was held fixed at 200 consistent with its setting during in-field image collection. Combined with the spectroradiometer responses, these values allow us to fully specify the sensitivity of each of the camera's three CMOS sensor channels to the entire visible spectrum.

2.3.2 - Estimating Camera Sensitivity

The camera's response to a monochromatic light will be proportional to the amount of light, the area of the aperture, and the exposure time. The constant of proportionality is considered the camera's sensitivity to that wavelength. For simplicity we normalize the sensitivity profiles to a peak of one. However, in principle this procedure could fully specify the camera's sensitivity on a per-quanta basis.

We begin with the illuminated test plate. The spectral irradiance $L_e(\lambda;\lambda_0)$ is a function of the wavelength parameterized by the wavelength setting on the monochromatic source. The units of this measure are in $W \cdot m^{-2} \cdot nm^{-1}$. Since the dial on the monochromatic light source may not be perfectly accurate, we define $\hat{\lambda}_0$ as our best estimate of the wavelength actually presented, i.e. the peak of the radiometric spectrum measured by the spectroradiometer. Since the light is not perfectly monochromatic, the irradiance (units $W \cdot m^{-2}$) is given by:

$$\hat{L}_e\left(\hat{\lambda}_0\right) = \int_{-\infty}^{\infty} L_e(\lambda;\lambda_0)\, d\lambda \ .$$

That is, the integrated spectral irradiance for a given monochromatic wavelength.

The camera sensitivities are a function of the wavelength, and the lights presented were not perfectly monochromatic. Thus, for a particular setting of the monochromator

technically the camera's response, e.g. to the red channel with sensitivity profile $S_R(\lambda)$, will be proportional to:

$$R\left(\hat{\lambda}_0\right) \propto \int_{-\infty}^{\infty} L_e(\lambda;\lambda_0) S_R(\lambda) d\lambda .$$

However, given that the pass band of the monochromator is relatively narrow compared to our step size, it is reasonable to estimate our sensitivities assuming a truly monochromatic light. That is, the response of the three channels will be given by:

$$\begin{aligned} R\left(\hat{\lambda}_0\right) &= K_R \hat{S}_R\left(\hat{\lambda}_0\right) \hat{L}_e\left(\hat{\lambda}_0\right) A\left(\hat{\lambda}_0\right) T\left(\hat{\lambda}_0\right) \\ G\left(\hat{\lambda}_0\right) &= K_G \hat{S}_G\left(\hat{\lambda}_0\right) \hat{L}_e\left(\hat{\lambda}_0\right) A\left(\hat{\lambda}_0\right) T\left(\hat{\lambda}_0\right), \\ B\left(\hat{\lambda}_0\right) &= K_B \hat{S}_B\left(\hat{\lambda}_0\right) \hat{L}_e\left(\hat{\lambda}_0\right) A\left(\hat{\lambda}_0\right) T\left(\hat{\lambda}_0\right) \end{aligned}$$

where $A\left(\hat{\lambda}_0\right) = \frac{a}{a_0}$ is the ratio of the aperture area a to the largest possible aperture area a_0, $T\left(\hat{\lambda}_0\right)$ is the shutter open time in seconds, while e.g. K_R is a normalizing constant for the red channel and $\hat{S}_R\left(\hat{\lambda}_0\right)$, which we take as a surrogate for $S_R(\lambda)$, is our estimate of the channel's sensitivity profile under the monochromatic source assumption. From here it is trivial to solve for the measured camera sensitivity profiles seen in Figure 9a.

2.3.3 - Mapping to Standard Color Spaces

The purpose of measuring the camera's spectral sensitivity is ultimately to find a simple mapping between CMOS responses and standard color spaces. Consider the XYZ standard established by the CIE. The standard is defined by a relationship between color coordinate X,Y,Z, and the associated color matching functions $\bar{x}(\lambda),\bar{y}(\lambda),\bar{z}(\lambda)$. Further, since this is a photometric rather than radiometric representation there is a

conversion from watts to lumens. For example, for the Y (luminance) channel the equation is $Y = 683lm/W \int \bar{y}(\lambda) L_e(\lambda) d\lambda$.

We would like to approximate these color matching functions with a triplet of linear weights on the camera sensitivities. For example, we would like a set of weights such that: $\bar{y}(\lambda) = w_{RY} S_R(\lambda) + w_{GY} S_G(\lambda) + w_{BY} S_B(\lambda)$. With these weights we have the following approximation:

$$\begin{bmatrix} \bar{x}(\lambda) \\ \bar{y}(\lambda) \\ \bar{z}(\lambda) \end{bmatrix} \approx \begin{bmatrix} w_{RX} & w_{GX} & w_{BX} \\ w_{RY} & w_{GY} & w_{BY} \\ w_{RZ} & w_{GZ} & w_{BZ} \end{bmatrix} \begin{bmatrix} S_R(\lambda) \\ S_G(\lambda) \\ S_B(\lambda) \end{bmatrix}$$

In order to determine the optimal weights, we minimize the unsigned difference of the left and right side of the above approximation via a least squares fit. By minimizing these nine terms we obtain the optimal weights to satisfy:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \frac{683 a_0}{aT} \begin{bmatrix} w_{RX} & w_{GX} & w_{BX} \\ w_{RY} & w_{GY} & w_{BY} \\ w_{RZ} & w_{GZ} & w_{BZ} \end{bmatrix} \begin{bmatrix} K_R^{-1} R \\ K_G^{-1} G \\ K_B^{-1} B \end{bmatrix}.$$

Similarly we can apply the same logic to the Stockman and Sharpe (Stockman, Sharpe et al. 2000) cone fundamentals. We use L, M, S , to represent the responses of long, medium, and short wavelength sensitive cones. Thus, we have:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \frac{683 a_0}{aT} \begin{bmatrix} w_{RL} & w_{GL} & w_{BL} \\ w_{RM} & w_{GM} & w_{BM} \\ w_{RS} & w_{GS} & w_{BS} \end{bmatrix} \begin{bmatrix} K_R^{-1} R \\ K_G^{-1} G \\ K_B^{-1} B \end{bmatrix}.$$

After minimizing the above equations, the following values were obtained for the weights:

$$\begin{bmatrix} w_{RX} & w_{GX} & w_{BX} \\ w_{RY} & w_{GY} & w_{BY} \\ w_{RZ} & w_{GZ} & w_{BZ} \end{bmatrix} = \begin{bmatrix} 1.230 & 0.643 & 0.042 \\ 0.167 & 0.909 & -0.209 \\ 0.137 & -0.145 & 1.655 \end{bmatrix}$$

$$\begin{bmatrix} w_{RL} & w_{GL} & w_{BL} \\ w_{RM} & w_{GM} & w_{BM} \\ w_{RS} & w_{GS} & w_{BS} \end{bmatrix} = \begin{bmatrix} 0.807 & 0.247 & 0.006 \\ 0.820 & 0.984 & -0.104 \\ -0.161 & -0.07 & 0.855 \end{bmatrix}.$$

Comparisons between the standard XYZ and LMS spaces and their respective fit transformations of the CMOS sensitivities can be seen in Figure 9b,c.

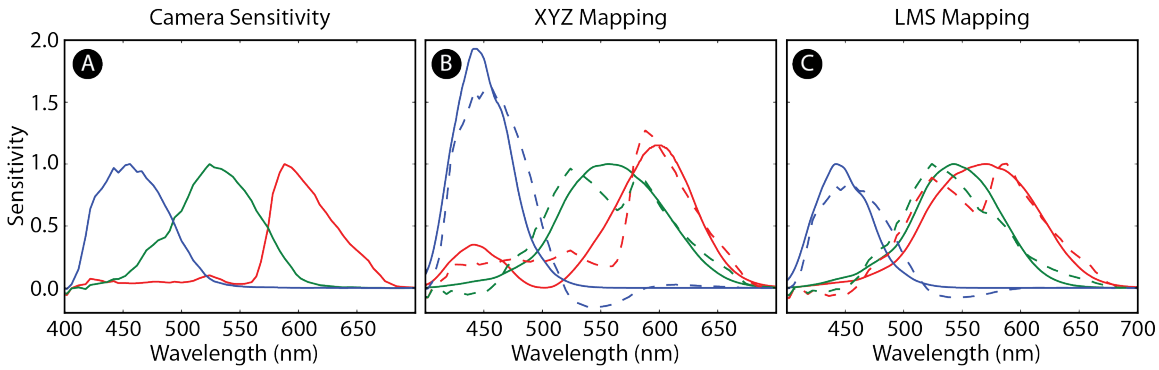

Figure 9 – Camera Spectral Sensitivity and Color Mapping

(A) The measured and normalized camera sensitivities. Note that these show the normalized sensitivities (i.e. accounting for K). (B) The best-fit mapping between the color matching functions of the XYZ color space (solid) and the mapped camera sensitivities (dashed). (C) The best-fit mapping between the Stockman and Sharpe (2000) long, medium, and short wavelength sensitive (LMS) cone fundamentals (solid) and the mapped camera sensitivities (dashed).

2.4 - STIMULUS GENERATION

The images that resulted from the above transformations were too large to be wieldy for further analysis or display. Further, the representation was inconvenient for thinking intuitively about the images. For example, the backwards slant of the camera implies that the camera's image plane was not perpendicular to the ground. This imaging geometry does not match our usual intuitions.

In keeping with the window analogy, I developed a piece of software that allowed the user to define a display in 3-space. Specifically, the user provides a 3xN matrix of xyz points (not to be confused with XYZ color space) corresponding with the locations of all N pixels in the display. These locations are then treated like any other xyz point, and mapped to the stereoscopic images and range images according to the calibration described in section 2.2. Thereby, both an RGB response, and xyz position can be defined for each pixel in the projection. For rectilinear displays, a useful analogy is a window screen. Rays from the scene are projected towards the viewer through the screen.

Clearly, the display pixel locations will not correspond exactly to locations sampled in the original data. Otherwise, there would be no call for resampling. Therefore, the mapping needed to be interpolated somehow. In actuality the rays will project to real valued pixel locations in images of the sort depicted in Figures 4-5 (usually tracing out a trapezoidal projection). The RGB data was interpolated with standard linear interpolation of the 16-bit images. The xyz data was interpolated using nearest neighbor (in order to preserve the sharpness of discontinuities.

The captured images were checked for object motion (i.e. visually inspected for troublesome regions of registration). Eighty images survived the filtering and became

stimuli in the final experiment. Thumbnails of the eighty images and corresponding data can be seen in Figures 10-11.

Figure 10 – Thumbnails of the Experimental Stimuli

Thumbnails of the images used in the psychophysical experiments. Stimuli used were a fraction of the size of the raw images because of the screen's limited field of view.

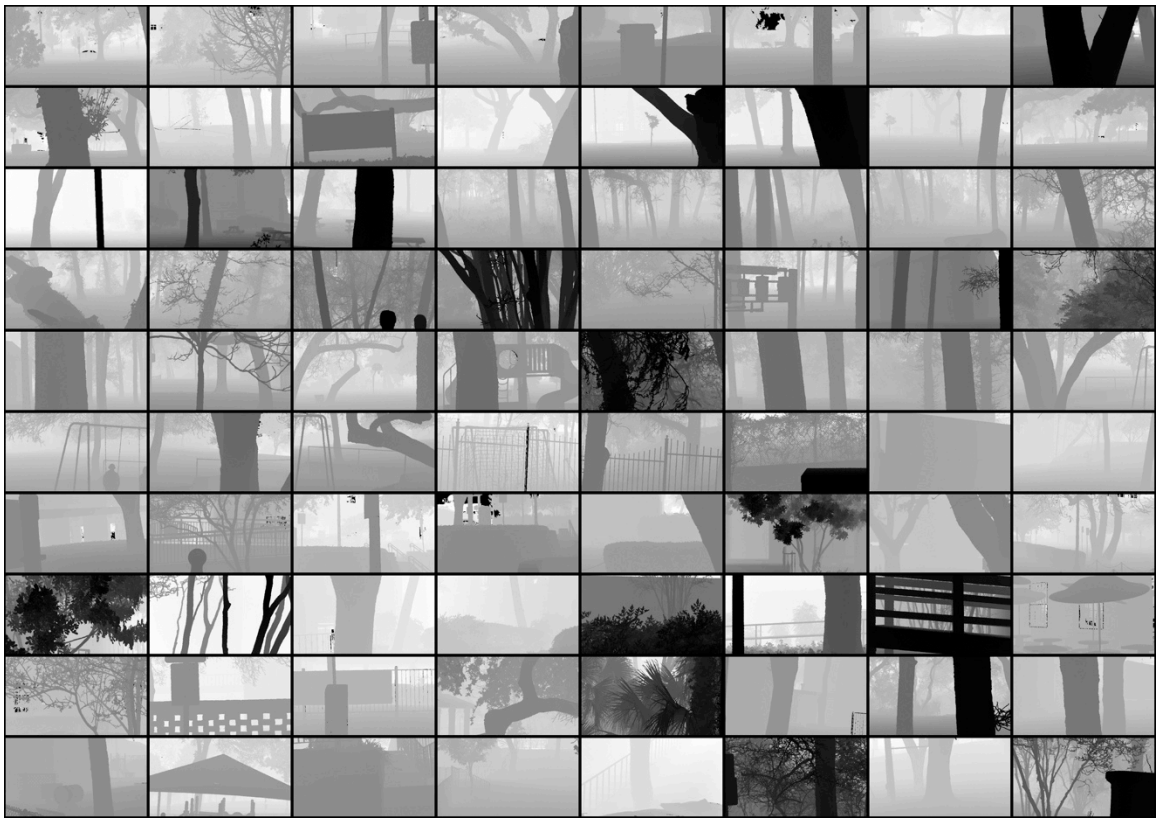

Figure 11 – Thumbnails of the Corresponding Range Images

Range images corresponding to the experimental stimuli in Figure 10. Measuring ground-truth allows for objective evaluation of performance.

Chapter 3: Psychophysical Methods

3.1 – Apparatus Overview

Depth acuity has never been measured in real scenes. It is logistically too hard. In laboratory experiments, the stimuli are deprived of the rich structure you see in real scenes. Nobody believes computer-generated images are of real scenes. Therefore they serve as a poor experimental model for natural acuity. The goal here is to create a naturalistic proxy for real scenes in the lab. Thus depth acuity measurements can finally be tractable to measure under real world conditions.

To this end a virtual reality display was designed which allowed me to present in the lab the stimuli measured in Chapter 2. The display was designed to appear as a window (an aperture) through which the scenes could be viewed stereoscopically. Since the screen was as near as the nearest object, all objects in the depicted scene could plausibly be behind the window. The room was dark, and the edges of the screen were obscured with black felt. The projected images were aligned carefully with the edges of the display completing the illusion (see Figure 12).

The screen itself (draper uniformity display) was tensioned in the frame. Thus the viewing geometry was not distorted by waves in the screen. The screen was rectangular, 1.43m wide and .80m tall. At the viewing distance of 3m this implies a visual angle of 27 degrees horizontally and 15 degrees vertically. The images were displayed at the standard 720p (1280x720px) definition. Therefore pixels subtended about .02 degrees, suggesting a Nyquist frequency of 25 cycles-per-degree (roughly half human limits). While higher resolution would be ideal, the sampling rate was matched to the camera.

The projector was a DepthQ HDs3D-1 stereoscopic display. Stereoscopy was achieved by active stereoscopy. That is, liquid crystal shutters (nVidia 3d Vision)

synchronized to the frame rate. The projector displayed frames at 120 Hz, alternating between the left and right eyes' views of the scene. Therefore, since the shutter occluded each eye on alternating frames, there was a 60Hz stereoscopic frame rate. At those temporal frequencies the flicker was undetectable. The remaining percept is a fused stereoscopic view of the scene.

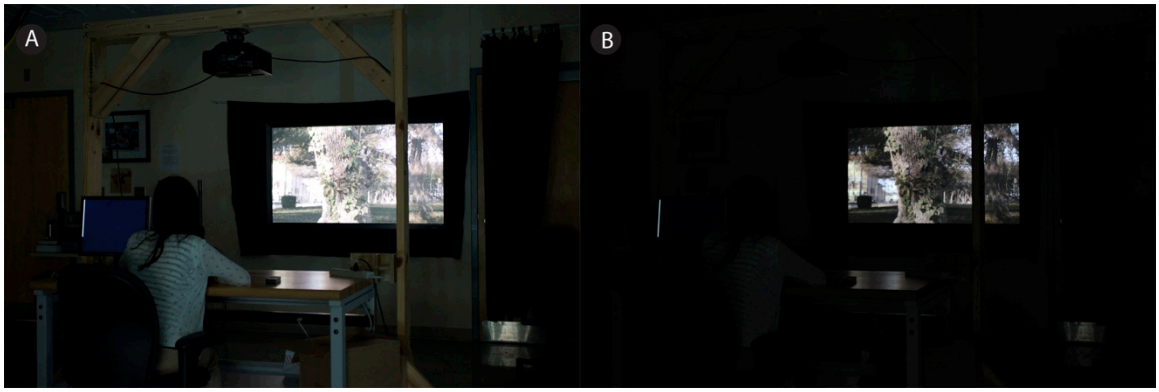

Figure 12 – Display Apparatus

(A) The camera gain is turned up so that the apparatus can be seen. The wooden gantry allows the display to be moved out of the way, necessary to accommodate other uses of the space. (B) The same view without the high camera gain. The black felt visible in (A) helps to remove cues to the edges of the frame.

Clearly, if the head moves, the world appears to distort. The image is correct from only one location. Therefore, a chinrest was used to minimize head movements. More precisely, the image is only correct from the location of the two eyes. It is distorted for observers with an IPD different than the 65mm used during scene measurement.

Accordingly, observers with commensurate IPDs were selected *post-hoc*. The viewing geometry was correct to a high precision. Consequently, objects appeared to be correct in size, the ground plane is slanted correctly, and so forth. Correct scale adds considerably to the subjective realism.

3.2 – PSYCHOPHYSICAL TASK

The task under consideration was a near-far distance discrimination task. On a given trial two points were indicated to the observer, one to the left and one to the right. The observer would press the right arrow if the right target was nearer, or the left arrow if the left target were nearer. Note that the points under consideration were points in the virtual scene. Subjects were instructed to consider the surfaces in the image. Since the distances between points in the scene are known, performance can be evaluated objectively.

The graphics were rendered in a custom OpenGL program written in the C++ programming language. The program was relatively simple, as the images were preloaded, and could for the most part be directly copied into memory. OpenGL was primarily used to display target indicators that identified points to discriminate during the task.

The optimal way to indicate the points is up for debate. Here we used a triangle pointing towards the center of a small circle denoting a surface. The point at the very center of the circle is the point being judged. In order to avoid introducing any additional stereoscopic information with the indicator, it was presented in the right eye's image only. Since the indicator was an abstract shape presented monocular it had no real apparent depth in the scene. The indicator appeared as needed in response to a press of the up arrow key. If the key was held down the indicators would be visible for 250ms

followed by 250ms of invisibility. Trials were self-timed and randomly interleaved. Fixation was uncontrolled. See Figure 13 for a schematic representation of the apparatus and the task.

For the monocular version of the task the left eye was patched for all subjects. Shutter glasses were still worn in case the glasses themselves degraded the image in any way. With head rested in the chinrest, and with the left eye patched, the illusion of a window was still convincing (all the monocular cues were correct given the position of the eye).

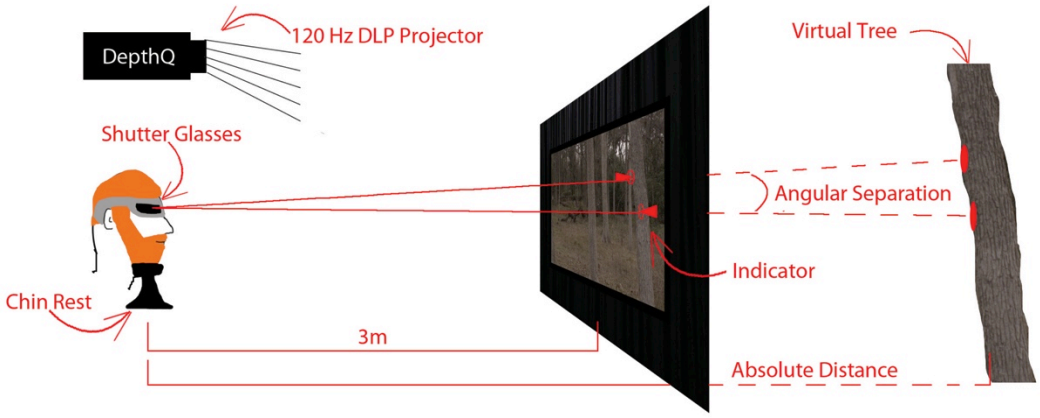

Figure 13 – Schematic of the Psychophysical Task

The observer judges the nearer of two points indicated in the virtual scene. Difficulty can be increased by the visual angle separating the points and the absolute distance of the virtual objects.

The purpose of the approach was to understand depth acuity in the way that we would understand other acuities. It was important to ensure that viewing conditions were chosen to cover the relevant space. Further, in order to create psychometric functions, the discriminations need to be sufficiently difficult. Therefore, targets could not be sampled completely randomly from scenes. Some effort needed to be made to ensure that there were both hard and easy trials. See Figure 13 for a schematic of the task. The next section goes into more detail on target sampling procedures.

3.3 – Sampling Procedure

The goal was to choose points as agnostically as possible. Still, enough difficult and easy trials needed to be selected to form a psychometric function. It is important to consider that while the mapping between range and image pixel value is known, it is not really under control. Stimuli cannot be designed, they must be found in the available scenes.

A number of factors may influence the difficulty or ease with which a particular discrimination could be made. In the spirit of agnostic sampling, RGB image properties were not considered. Specifically, image-based parameters, notably contrast, were not systematically sampled to vary difficultly. Instead, difficulty was assigned according to three metrics, (i) the difference in distance between the two points, (ii) the absolute distance to the points, and (iii) the visual angle separating the points.

The first metric is the natural measure of difficulty, the similarity in distance between the two points. Clearly, the more similar in distance, the more difficult the discrimination must be. Since the only information available to anchor depth acuity is from the disparity literature, fixed disparity bins were used to select points. This turned

out to not correspond with behavior (see Chapter 4 for results), and so additional trials were added to help subsequent subjects reach threshold.

The second difficulty metric was the absolute distance to the judgment in question. The absolute distance to the targets will impact performance both because of perspective projection, and the degradation of stereoscopic cues to depth. In the study distances were sampled from 3m-45m. Sampling was spread out to cover the distances of interest, with bins becoming larger (sampling sparser) at larger distances. An effort was made to spread trials uniformly across images, although the prevalence of content at a given distance inevitably biased the sampling towards a subset of scenes in certain conditions.

The third difficulty metric is the visual angle separating the two points. It is known that sampling in the retina can account for the drop in performance in fine acuity tasks in the periphery(e.g., Arnow and Geisler 1996). Further, disparity detection thresholds are known to vary with the eccentricity of the target (Blakemore 1970). Thus it is reasonable to assume that the visual angle separating the two targets will have an impact on performance. Fixation is uncontrolled, so it is not possible to describe viewing conditions perfectly. However, the solid angle separating the targets naturally constrains the observer's ability to have simultaneous foveal views of the two targets. Sampling along the visual angle difficulty axis was concentrated around 2°, 5°, and 10° of separation. Again, attempts were made to uniformly sample the scenes, but the relative locations of objects in the scenes ultimately constrained the success of these attempts.

In order to find each sample, a random point in a random scene was chosen on account of its distance. This was taken as the more distant of the two points because crossed disparities are easier to find than uncrossed disparities at large distances. The

second point was chosen on account of the difficulty needed. First, an annulus of pixels corresponding to the appropriate angular separation bin was selected. These pixels were checked for delta-scene-distances appropriate for the needed difficulty level (commensurate with a given disparity along primary gaze). If the regions surrounding the points were approximately planar (verified via Singular Value Decomposition), the pair was accepted. One final check was performed to ensure that the targets were not occluded in either eye. Once a pair of targets was selected, the neighboring region was ruled out and the process began again until all needed trial bins were filled.

Monocular trials were still sampled according to 'disparity' even though it is a meaningless concept. Ultimately, dioptric difference rather than disparity was adopted as the primary difficulty dimension because it had a closer (one-to-one) relationship with absolute difference in distance. However, these measures of difference in distance are closely related, and the dioptric difference can be thought of as a surrogate for disparity.

No visual estimations were repeated by the same observer, however two groups (pairs) of observers received equivalent trials. Chapter 4 goes into more detail on how observer identity (individual differences) and stimulus particulars impact performance.

Chapter 4: Experimental Results

4.1 – Psychometric Functions

It turned out that binning on either disparity or dioptric differences was not optimal, because performance turned out to be better described by thresholds that increased in proportion to absolute distance (i.e. constant Weber fraction). Consequently, the sampling was not ideal, and at near distances (where thresholds were highest) observers often had difficulty reaching 100% correct. Nonetheless, the psychometric functions in most cases appeared good. Example psychometric functions for one subject can be seen in Figure 14.

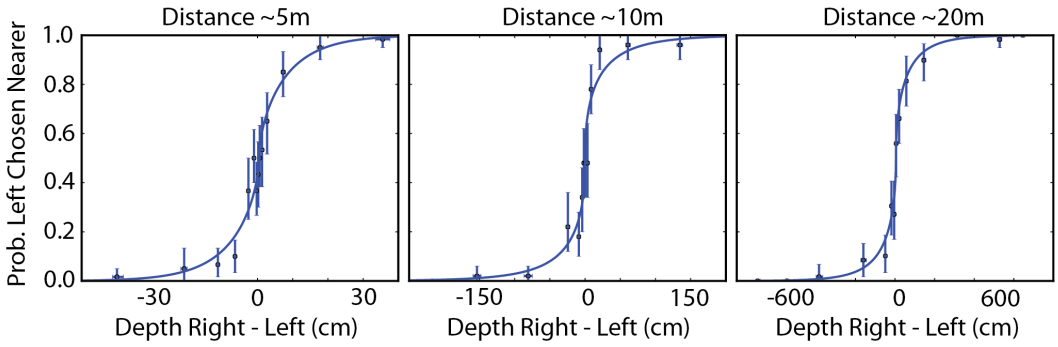

Figure 14 – Example Psychometric Functions

Psychometric functions fit to some of the data by subject SVV. Note that threshold increases (x axis change) as measurements are made at increasing distances. All of these measurements were for a visual angle separation of around 5°. Curves represent maximum likelihood fits to the data.

Since difficulty varied along three dimensions, it is impossible to perfectly isolate the influence of all three. Usually, visual angles were binned into three bins 0°-2°, 4°- 6°, and 8°-12°. Distances were binned more continuously, with bin widths being roughly proportional to the distance. That is, a psychometric function describing performance at a distance of '10m' would include trials anywhere from 5m-15m. The primary difficultly axis was sampled in many different ways depending on convenience, and maximum likelihood methods were used to fit the psychometric functions using all of the relevant trials respecting their continuous location on the x-axis. In the plots, boxes show binned data with horizontal and vertical error bars. They provide a visual aide for assessing the quality of the fit, but the binned percent corrects were not used in the model fit.

4.2 – Observer Agreement

Using the technique described in the previous section, psychometric thresholds could be obtained for each observer at the various distances. Displayed in Figure 15 is threshold (measured as a Weber fraction in percent) of the four observers in the experiment. In the figure, each dot corresponds to a threshold that came from a psychometric function such as those depicted in Figure 14. Note considerable overlap between the green and purple observers. They were tested using matched trials. In all, performance seemed roughly consistent across observers. Figure 15 provides some justification for aggregating thresholds across observer judgments in order to more densely sample the space.

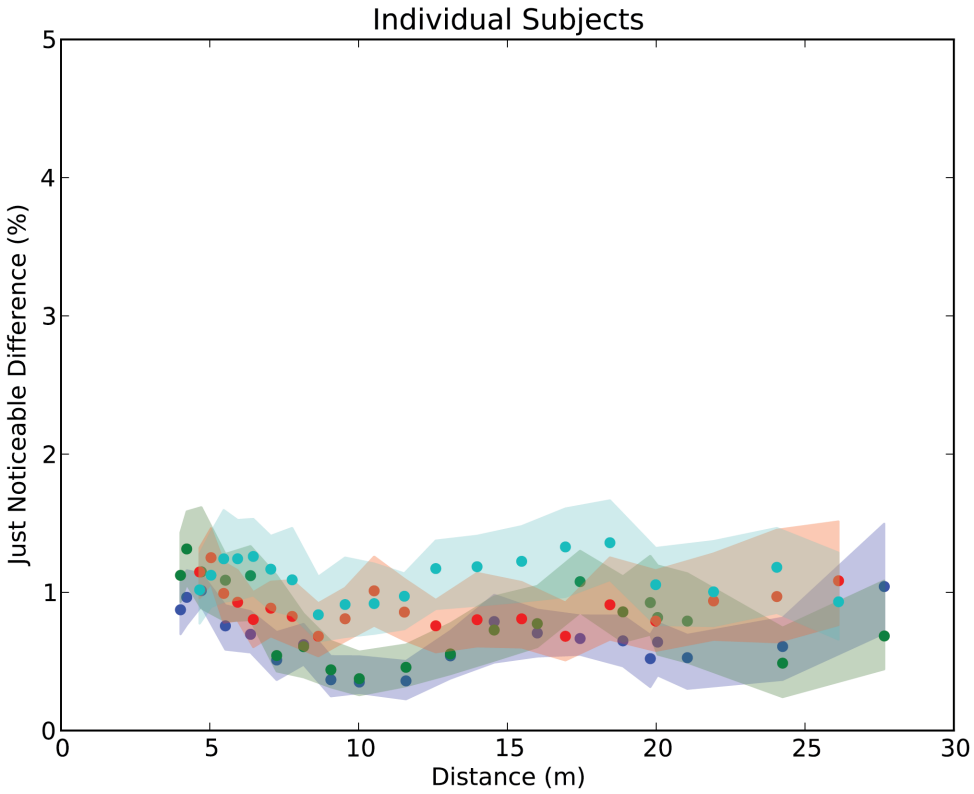

Figure 15 – Performance of Individual Subjects

The plot shows thresholds for individual subjects (colors) in the near far task. The x-axis indicates the distance of the discriminations in the psychometric functions. The y-axis shows the just noticeable difference in percent range (Weber Fraction Threshold). Dots show the measured threshold, with bootstrapped 95% confidence intervals shaded. All observations were at about 2°. Note that the observers largely fall on top of each other.

4.3 – BASIC ACUITY DESCRIPTION

Two possible predictions were visually compared, constant fractional performance in absolute distance, and constant disparity performance. Visually speaking, it appears that constant fractional (Weber) performance is the better description of the data (see section 4.4 for disparity performance). Further, the observers' performance was strongly affected by the visual angle separating the targets. Figure 16 shows the overall description of average observer performance as function of absolute distance and target separation. In effect, Figure 16 completes my second aim. It suggests that distance discrimination threshold is approximately a fixed fraction of the absolute distance, with a fraction that increases with separation angle.

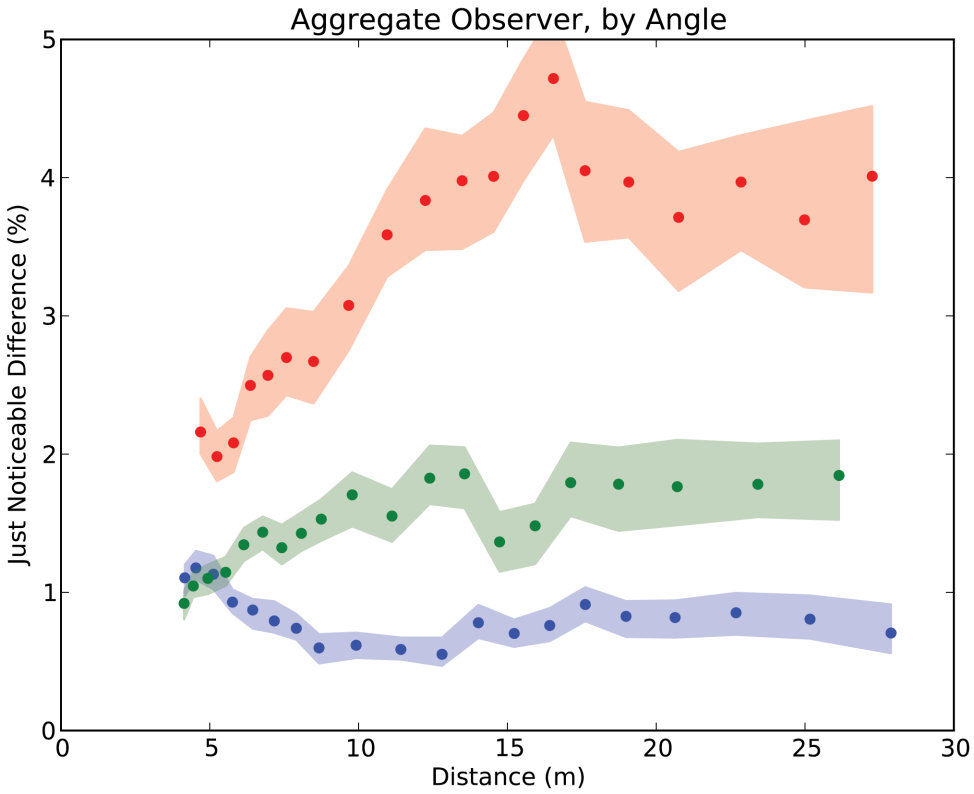

Figure 16 – Acuity Description of the Aggregate Observer

The figure shows the description of the aggregated observer's overall performance. Blue shows the performance around 2° of angular separation. It shows the aggregation of the data in Figure 15 (axes are the same). The other colors depict other angular separations, i.e. green ~5°, and red ~10°. Shaded regions are bootstrapped 95% confidence intervals on thresholds. The impact of visual angle on threshold is substantial.

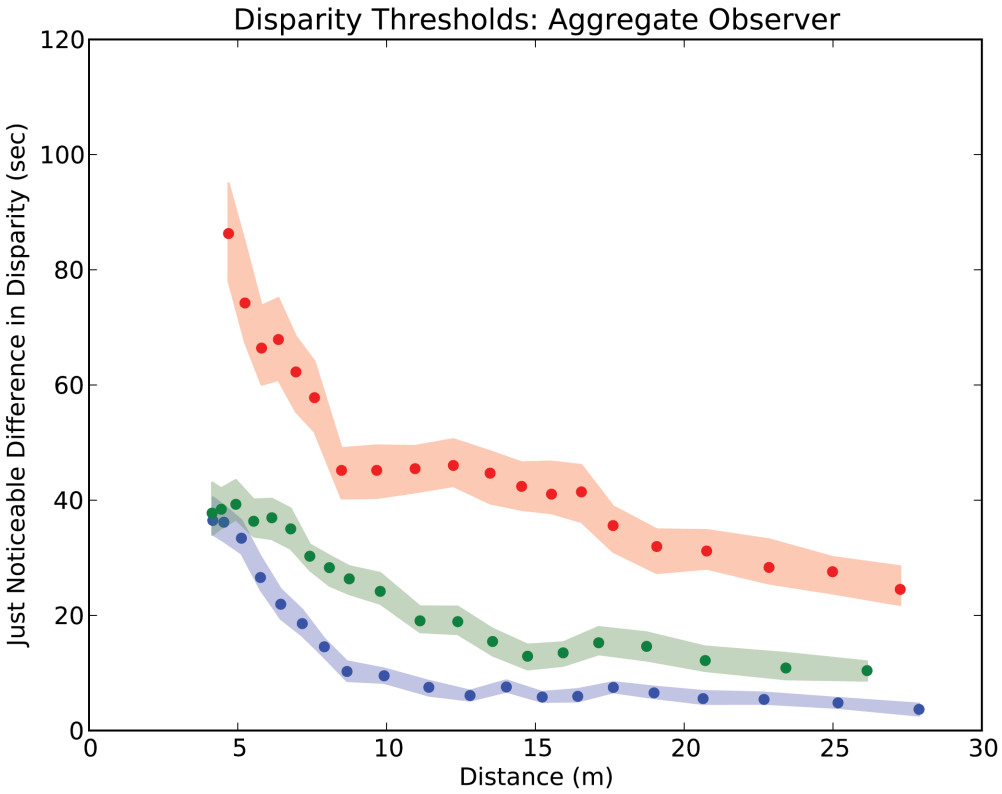

Figure 17 – Acuity Measured in Disparity

The analog of Figure 16. The only conceptual difference between the figures is that threshold (y-axis) is now measured in arc seconds of disparity rather than percent distances. Note that thresholds monotonically decline by this measure. The prediction of the 'disparity strategy' is that thresholds in disparity would be constant as a function of distance.

4.4 – Performance Measured As Binocular Disparity

It can be convenient to represent performance as thresholds in binocular disparity. This can help relate these acuity measurements to the disparity literature. When measured as a disparity, thresholds clearly decrease as a function of distance. This should not be taken as evidence that threshold disparity is improving. Rather, monocular cues are likely playing a greater role with greater distance, improving the quality of the estimate. That is, it is important to remember that the units describing the estimates precision have nothing to do with how the estimate was generated. Figure 17 shows the analog for Figure 16 with arc seconds of disparity on the y-axis.

4.5 – Monocular Comparison

Recall that depth discrimination performance was also measured with one eye patched. These conditions were identical to the baseline binocular case except for the availability of the binocular disparity cue, thus they show the performance based purely on the monocular cues. The red data points in Figure 18 show depth discrimination as function of distance for the monocular cues. For comparison, the blue points are the depth discrimination data for the binocular case in Figure 16. It is clear that thresholds are elevated in the monocular condition, especially at near distances. From the plot it appears that binocular and monocular performance merge at a distance of around 15m (see Chapter 5).

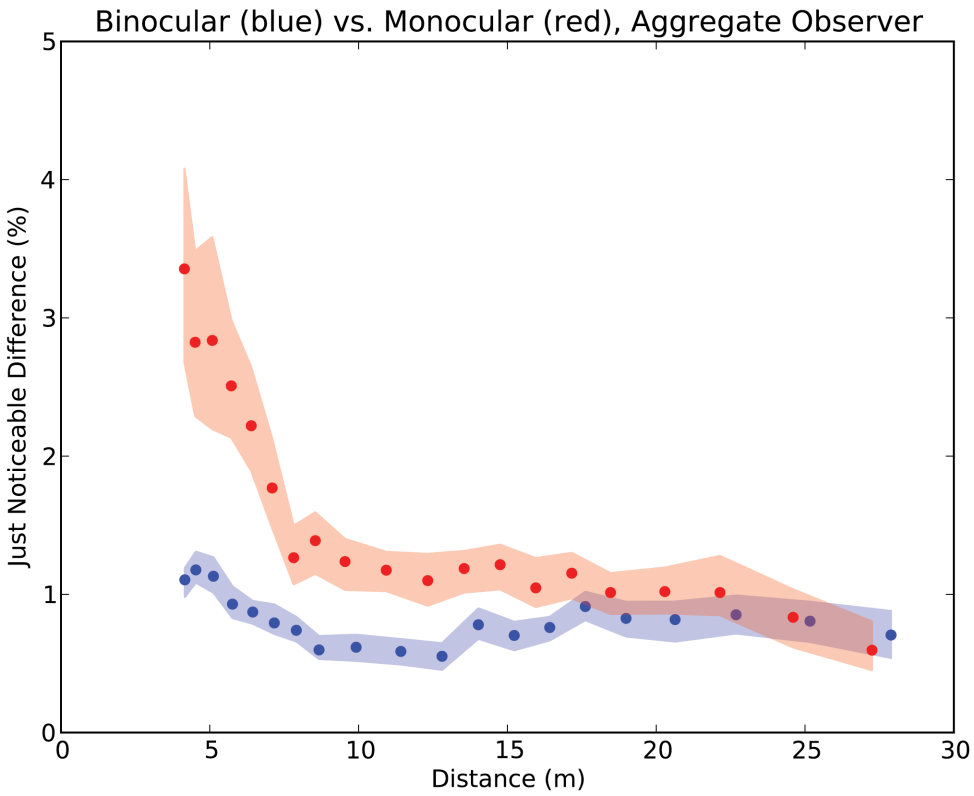

Figure 18 – Foveal Binocular and Monocular Comparison

The blue data is the exact same as the 2° data from Figure 16. The red curve shows the aggregated monocular comparisons for the same four subjects (different target locations). Shaded regions denote 95% bootstrapped confidence intervals. Monocular thresholds are clearly elevated up to a distance of about 15m. Thus, binocular disparity is likely contributing to those judgments.

Another useful way to look at the monocular thresholds is, ironically, in units of binocular disparity. Specifically, Figure 19 plots are the monocular thresholds in disparity units, as if the eye had not been patched. It is a bit counter-intuitive to measure

performance in this way, but it is useful for comparison with binocular conditions. In other words, Figure 19 shows the monocular analog to Figure 17. Note the substantial elevation in thresholds for near judgments compared to those in Figure 17.

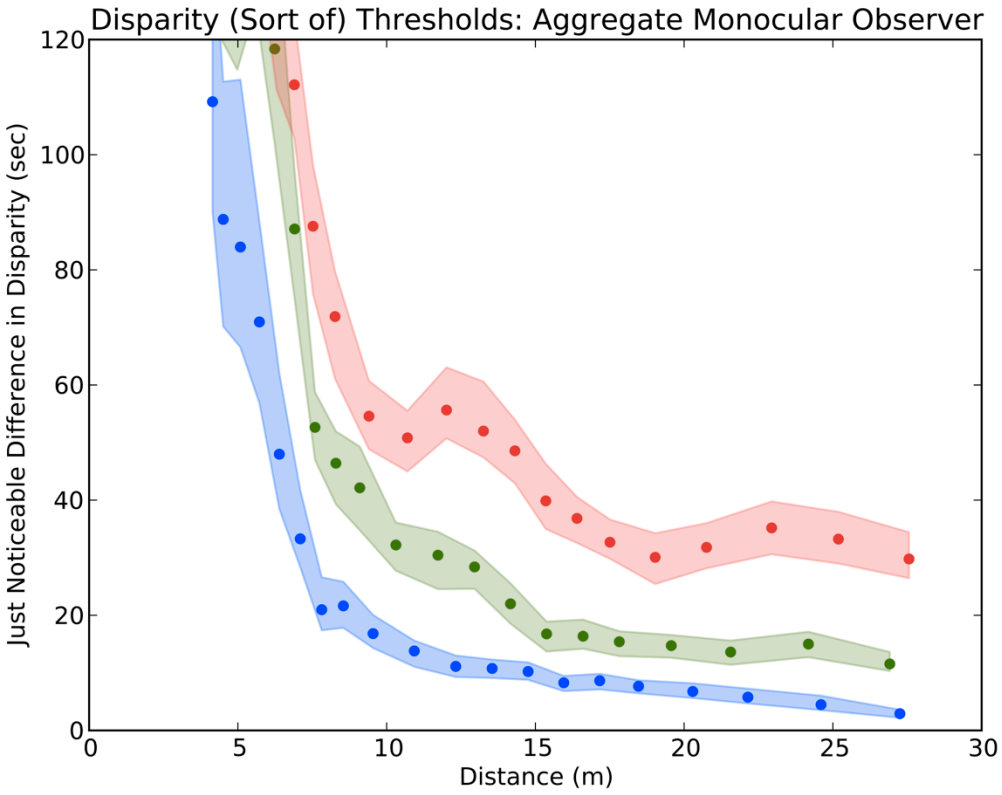

Figure 19 – Monocular Thresholds Expressed as Disparity

Thresholds measured as disparities in the monocular condition. Axes and color codes were kept the same for comparison with the binocular analog Figure 17. Thresholds off the chart in the 5° and 10° are so elevated the plot becomes difficult to interpret if the axes are scaled. Again, around 15m thresholds resemble those measured in the binocular condition.

4.6 – CUE COMBINATION PREDICTIONS

The variability associated with the sensory estimates measured in the two experiments can be combined to create a prediction for a hypothetical isolated disparity condition. Recall that under the assumption of optimal combination of independent cues the variance of the combined estimate is $\sigma(d,\theta)^2 = \frac{1}{\sigma_m(d,\theta)^{-2} + \sigma_b(d,\theta)^{-2}}$. In the experimental context, the variability associated with the estimate is directly related to the threshold in the combined (binocular) condition. It is assumed that there are two cues with associated standard deviation parameters, disparity σ_B, and monocular σ_M. Therefore, by reciprocating both sides we see that the reliabilities add, $\sigma(d,\theta)^{-2} = \sigma_m(d,\theta)^{-2} + \sigma_b(d,\theta)^{-2}$. Exploiting this, simple rearrangement yields predictions for thresholds in the supposed disparity condition, i.e. $\sigma_b(d,\theta) = \left(\sqrt{\sigma_m(d,\theta)^{-2} - \sigma(d,\theta)^{-2}}\right)^{-1}$. Note that the right hand side of the equation is completely constrained by the experimental thresholds. The dots in Figure 20 show the parameter-free predicted disparity thresholds from the current study with the above formula.

Blakemore (1970) measured disparity thresholds for two observers as a function of distance from the horopter and visual angle eccentricity. While it is impossible to know where subjects were fixated in the current task, it is plausible that they fixated back and forth between targets. Therefore, the relative horizontal disparity would reflect the disparity on the horopter, at the eccentricity defined by their visual angle separation. Accordingly, the colored bands in Figure 20 show the range of disparity detection thresholds (from the two observers) measured in Blakemore's experiment for the foveal (blue), 5° (green) and 10° (red) conditions. As can be seen, the thresholds predicted from the current data largely fall within the range expected from the Blakemore study

(although there are a few outliers). This shows, rather remarkably, that the effect of the binocular cues under natural viewing conditions is largely consistent with measurements of disparity thresholds with simple laboratory stimuli.

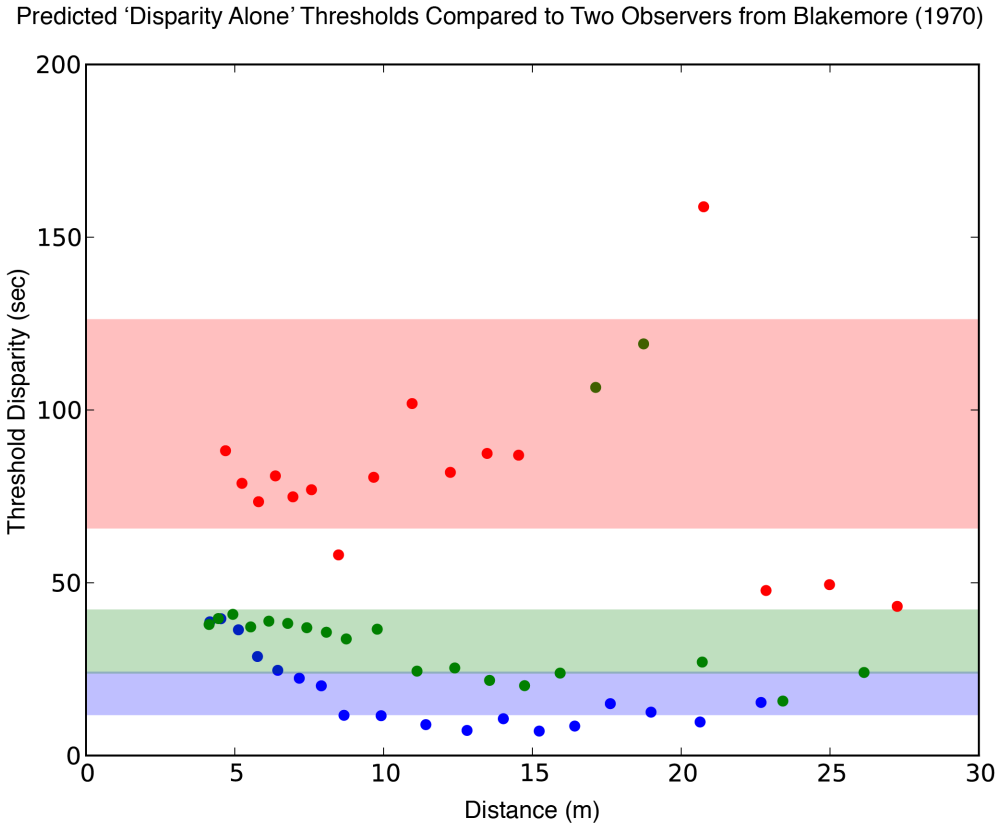

Figure 20 – Disparity Threshold Predictions Compared to Blakemore

Dots show predicted thresholds in a disparity alone condition. Shaded regions cover the range defined by the disparity detection thresholds of two observers in Blakemore (1970). Colors denote an eccentricity (or spatial separation) of foveal (blue), 5° (green), and 10° (red). Aside from a few scattered points, dots fall near the isolated disparity thresholds

from the literature. Disparity threshold predictions appear to decrease slightly with distance.

From the above plot it is not entirely obviously just how remarkable these predictions are. In particular, it is difficult to appreciate how sensitive the predictions are to the original data. For example, if threshold in the monocular case happens (by chance) to be even a second of arc lower than in the binocular case than the prediction has no real-valued solution. Perhaps a more intuitive way to see the accuracy of these predictions is to pretend the experiment had been of a more standard cue-combination design. That is, we can sum the reliability implied by Blakemore's thresholds with the reliabilities implied by the measured monocular thresholds. A simple transformation of the sum of these reliabilities produces a prediction for threshold in the combined (binocular viewing) condition. The shaded regions and filled marks in Figure 21 depict the same binocular thresholds as Figure 17. The open symbols and dashed lines depict the monocular thresholds seen in Figure 19. Combining these monocular thresholds with Blakemore's disparity thresholds yields a prediction (solid line). Note the agreement. Again, this is a zero free-parameter prediction.

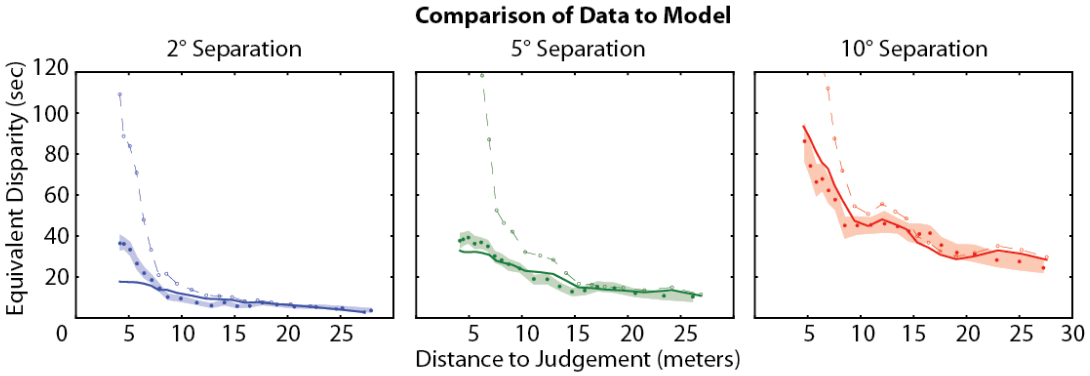

Figure 21 – Predicted Thresholds Comparison

The filled dots and shaded regions depict the binocular thresholds. The open dots and dashed lines depict monocular thresholds. Solid lines show the prediction yielded by combining the monocular data with Blakemore's disparity thresholds.

Oruç et al. (2003) derived the optimal combination rule for possibly correlated cues. Borrowing their equation 7 the optimal combination rule for this task would result in threshold predictions as follows:

$$\sigma(d,\theta)^{-2} = \frac{\sigma_m(d,\theta)^{-2} + \sigma_b(d,\theta)^{-2} - 2\rho\sqrt{\sigma_m(d,\theta)^{-2}\,\sigma_b(d,\theta)^{-2}}}{1-\rho^2}.$$

To test the importance of correlations I repeated my predictions assuming a (decidedly generous) correlation coefficient $\rho = 0.3$. These predictions can be seen in Figure 22, analogous to Figure 21 with only the prediction lines changed. As can be seen from the plot, the predictions are qualitatively unchanged, although are made a bit worse by the addition of an assumed correlation. Notice that one change in the prediction not

supported by the data is that the binocular and monocular performance is predicted to converge at a nearer distance (closer to 10m).

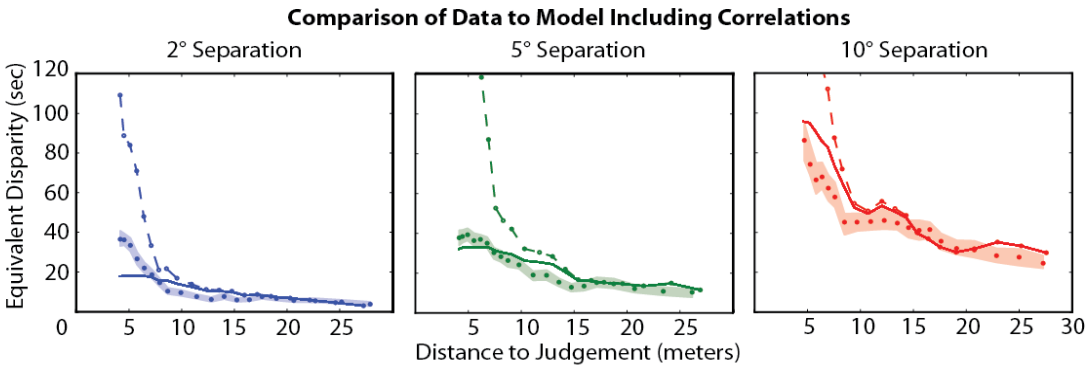

Figure 22 – Predicted Thresholds Comparison Including Correlations

Analogous plot to figure 21. Here the solid lines reflect predictions using Blakemore's data, and an assumed correlation of 0.3.

Chapter 5: Discussion

5.1 – Natural Scenes Database

The natural scenes database collected here should prove valuable to the field far beyond the scope of this work. One major conclusion of this work is that natural scenes are indeed regular enough for study. In fact, depth perception depends on exactly this same regularity. If the real world were not highly statistically regular, then the visual system could not exploit those regularities to do the task. Certainly, pictorial cues to depth would be useless. They all depend on assumptions about the structure of the world.

The psychophysical results reported here suggest that binocular disparity is the primary binocular cue to depth. However, monocular cues to depth have not been addressed in great detail. There is substantial work to be done teasing apart the relative importance of monocular cues in real scenes. As mentioned, monocular cues are highly dependent on regularities in the natural environment. Therefore, it is likely that there is substantial insight to be gained by studying the environment itself, i.e. the generative model of the stimulus.

The value of studying the statistics of natural scenes is well documented (Geisler and Diehl 2002). We have already begun the process of evaluating the importance of depth cues directly by statistical analysis of these scenes (Burge et al. in preparation). Preliminary results are promising, and show some counter-intuitive results, thus demonstrating the value of directly studying the availability of information in real scenes rather than relying strictly on scientific intuitions. Since this is the first dataset to rigorously couple the 3d structure of scenes with their stereoscopic projections it should have a substantial impact on the field.

5.2 – Depth Acuity

The depth acuity measurements were made using naturalistic stimuli, and therefore serve as an accurate representation of human uncertainty about distance across a large sample of trials in the real world. There are some caveats to this claim. One major consideration is that these are depth acuity measurements at largish distances, through a window (an aperture) with a stationary viewing position. That said, the measurements here are much more applicable to real world depth perception than previous depth discrimination measurements. These caveats are discussed more in section 5.4.

One finding, if unsurprising, is that depth discrimination performance does not track the performance of a 'disparity alone' strategy in real scenes. When measured as a disparity, thresholds drop precipitously as a function of distance. At least for foveal measurements (blue symbols in Figure 16), performance roughly follows Weber's law with a Weber fraction at about 1% of the absolute distance (thus, for foveal comparisons at 10 m, threshold will be ~1cm). This is consistent with increasing reliance on monocular cues to depth.

Another (unsurprising) finding is that increasing the angular separation between the points increases thresholds. The source of this increase in thresholds is less obvious. Two factors certainly play a role. First, disparity thresholds increase for more eccentric targets (Blakemore 1970). Second, planar (monocular) acuity changes as a function of the visual field (Robson and Graham 1981). The successful predictions by the cue combination model suggest it might be possible to untangle their relative contributions. The results of the cue combination modeling are discussed more in section 5.3.

In summary, this work represents perhaps the most comprehensive study of depth acuity in real scenes. The results are plausible, systematic, and relatively consistent across observers despite ambivalence towards image content.

5.3 – Cue Combination

The results of the cue combination modeling instilled confidence that my measurements are meaningful. Monocular thresholds were measurably higher than binocular thresholds at distances under 15m. This elevation in thresholds reveals the importance of binocular vision.

More impressively, the standard cue combination model predicted the effect of disparity on depth thresholds with high accuracy and no free parameters (Figure 21).

Blakemore's (1970) measurements have been relied upon for decades as an independent measure of disparity sensitivity. Turning the analysis around, the current results demonstrate that a somewhat accurate prediction of his isolated thresholds could be obtained by inference from independent measures in the full-cue, and monocular only conditions. A quick check to see the relevance of possible correlations between the cues shows that if the optimal rule is still used the predictions are largely unaffected, although near zero values for the correlation still appear to provide the best fit suggesting that the cues are in fact independent.

An additional benefit of the current approach is that Blakemore's measurements could be extended into distance. Measurements here suggest that disparity thresholds might actually drop somewhat as a function of distance. Despite disparity being an angular measurement, information quality in the image might change reliably with distance and therefore influence our angular sensitivity. Glennerster and McKee (1999) have shown that nearby references can improve disparity acuity. Liu et al. (2008) showed that uncrossed disparities are quite rare at large distances with an over-representation of zero disparities. Therefore, it is possible that references are useful and prevalent for distal comparisons.

Against the suggestion of Maloney and Landy (1989), this prediction is made without holding experimental ancillary measures fixed. Stimuli are free to vary as they do in real scenes. Thankfully, real scenes are drawn from a relatively stationary distribution. Reversing the prediction, i.e. predicting full-cue thresholds from monocular thresholds and Blakemore's disparity thresholds, reveals just how precise the predictions need to be in order to be this successful. Recall, there was zero control over the content of images.

Therefore, average performance is fortunately well-behaved over a relatively short, 'long-run.'

5.4-CAVEATS

One poor prediction in the data happens with foveal comparisons at the nearest distances. It is likely that defocus conflicts could play a role. Sebastian et al. (in press) measured defocus thresholds as low as 1/8 of a diopter in real scenes. Therefore, presented images were in good focus at 3m, and in progressively worse focus up to a distance of approximately 8m (where it would level off). It is likely that robust fusion mechanisms would take over once the conflict was severe enough, thus 'turning-off' defocus as a cue to depth. However, at these near distances, defocus could plausibly be used by the system, and therefore bias monocular depth estimates. This is worthy of further exploration.

Another coincidence is the convergence of monocular and binocular cues to depth at a distance of around 15m. It should be noted that these measurements were made through an aperture (the window). Therefore, the ground plane (an important cue to distance) was occluded by the wall out to some distance. It so happened that in this configuration, that distance was 15m. Since the absence of a visible ground plane caused an elevation of thresholds, it is worth changing the size of the aperture to investigate its significance.

Finally, heads were fixed. Clearly, head-free viewing has the potential to improve sensitivity. A long-term goal is to extend my psychophysical (and scene collection) methods to include parallax images over some reasonable ego motion (e.g. seated head movements). Doing so could improve the generality of the results.

5.5-CONCLUSIONS

In conclusion I have developed a large dataset of stereoscopic natural scenes pairing images with distance measurements at each pixel. From these images I created a naturalistic proxy for the real scenes by stereoscopically projecting the images on a simulated window. Using this naturalistic proxy, acuity for distance in real scenes was assessed over a wide range of viewing conditions. I found that to first approximation, thresholds followed Weber's law with a threshold of 1% of the distance the judgment was made at. Repeating the measurements under patched-eye viewing conditions directly assessed monocular thresholds. By comparing monocular and binocular performance an indirect measurement of the influence of disparity can be made. I found that these indirect measurements closely matched measurements from the literature, demonstrating the validity of the natural scene approach taken here.

References

Adelson, E. (1995). "Checker Shadow Illusion."

Allison, R., B. Gillam and E. Vecellio (2009). "Binocular depth discrimination and estimation beyond interaction space." Journal of Vision **9**: 1-14.

Ames Jr, A. (1951). "Visual Perception and Rotating Trapezoidal Window." Psychological Monographs: General and Applied **65**.

Arnow, T. and W. S. Geisler (1996). "Visual detection following retinal damage: predictions of an inhomogeneous retino-cortical model." Photonics West, International Society for Optics and Photonics.

Badcock, D. R. and C. M. Schor (1985). "Depth-increment detection function for individual spatial channels." J Opt Soc Am A **2**(7): 1211-1216.

Blakemore, C. (1970). "The range and scope of binocular depth discrimination in man." J Physiol **211**(3): 599-622.

Clark, J. J. and A. L. Yuille (1990). Data fusion for sensory information processing systems. Boston, Kluwer Academic Publishers.

Cormack, L. K., S. B. Stevenson and D. D. Landers (1997). "Interactions of spatial frequency and unequal monocular contrasts in stereopsis." Perception **26**(9): 1121-1136.

Cormack, L. K., S. B. Stevenson and C. M. Schor (1991). "Interocular correlation, luminance contrast and cyclopean processing." Vision Res **31**(12): 2195-2207.

Durgin, F. H., A. Hajnal, Z. Li, N. Tonge and A. Stigliani (2010). "Palm boards are not action measures: an alternative to the two-systems theory of geographical slant perception." Acta psychologica **134**: 182-197.

Durgin, F. H. and Z. Li (2011). "Perceptual scale expansion: an efficient angular coding strategy for locomotor space." Attention, perception & psychophysics **73**: 1856-1870.

Edgerton, S. Y. (1991). The heritage of Giotto's geometry : art and science on the eve of the scientific revolution. Ithaca, Cornell University Press.

Ernst, M. O. and M. S. Banks (2002). "Humans integrate visual and haptic information in a statistically optimal fashion." Nature **415**(6870): 429-433.

Felsen, G., & Dan, Y. (2005). "A natural approach to studying vision." Nature neuroscience, **8**(12): 1643-1646.

Fine, B. J. and J. L. Kobrick (1983). "Individual Differences in Distance Estimation: Compariosn of Judgements in the Field with those from Projected Slices of the Same Scenes." Perceptual and Motor Skills: 3-14.

Frisby, J. P. and J. E. Mayhew (1978). "Contrast sensitivity function for stereopsis." Perception **7**(4): 423-429.

Fukusima, S. S., J. M. Loomis and J. a. Da Silva (1997). "Visual perception of egocentric distance as assessed by triangulation." Journal of experimental psychology. Human perception and performance **23**: 86-100.

Geisler, W. S. and R. L. Diehl (2002). "Bayesian natural selection and the evolution of perceptual systems." Philos Trans R Soc Lond B Biol Sci **357**(1420): 419-448.

Gibson, E. J. and R. Bergman (1954). "The effect of training on absolute estimation of distance over the ground." Journal of experimental psychology **48**: 473-482.
Gibson, E. J., R. Bergman and J. PURDY (1954). "The effect of prior training with a scale of distance on absolute and relative judgments of distance over ground."
Glennerster, A. and S. P. McKee (1999). "Bias and sensitivity of stereo judgements in the presence of a slanted reference plane." Vision Res **39**(18): 3057-3069.
Gogel, W. (1972). "SCALAR PERCEPTIONS WITH Binocular Cues of Distance." The American journal of psychology.
Gordon, C. C., R. A. Walker, I. Tebbetts, J. T. McConville, B. Bradtmiller, C. E. Clauser and T. Churchill (1989). "1988 Anthropometric Survey of US Army Personnel-Methods and Summary Statistics.".
Green, D. M. and J. A. Swets (1966). Signal detection theory and psychophysics. New York,, Wiley.
Greenwald, H. S., & Knill, D. C. (2009). "A comparison of visuomotor cue integration strategies for object placement and prehension." Visual neuroscience, **26**(01): 63-72.
Hayhoe, M., B. Gillam, K. Chajka and E. Vecellio (2009). "The role of binocular vision in walking." Visual neuroscience **26**: 73-80.
Held, R. T., E. A. Cooper and M. S. Banks (2012). "Blur and disparity are complementary cues to depth." Curr Biol **22**(5): 426-431.
Holway, A. H. and E. G. Boring (1941). "Determinants of Apparent Visual Size with Distance Variant." The American Journal of Psychology **54**: 21.
Howard, H. (1919). "A Test for the Judgment of Distance." Am. J Ophthal **2**: 656-675.
Howard, I. P. and B. J. Rogers (2012). Perceiving in depth. New York, Oxford University Press.
Julesz, B. (1960). "Binocular Depth Perception of Computer Generated Patterns." Bell System Technical Journal.
Knill, D. C. (1998). "Discrimination of planar surface slant from texture: human and ideal observers compared." Vision Res **38**(11): 1683-1711.
Knill, D. C. (2007). "Learning Bayesian priors for depth perception." **7**: 1-20.
Knill, D. C. and W. Richards (1996). Perception as Bayesian inference. Cambridge, U.K. ; New York, Cambridge University Press.
Knill, D. C. and J. A. Saunders (2003). "Do humans optimally integrate stereo and texture information for judgments of surface slant?" Vision Res **43**(24): 2539-2558.
Lambert, J. H. (1760). I.H. Lambert Academiae Scientiarum Electoralis Boicae ... Photometria, siue, De mensura et gradibus luminis, colorum et umbrae. Augustae Vindelicorum, Sumptibus Viduae Eberhardi Klett, typis Christophori Petri Detleffsen.
Lee, B. and B. Rogers (1997). "Disparity modulation sensitivity for narrow-band-filtered stereograms." Vision Res **37**(13): 1769-1777.
Liu, Y., A. Bovik and L. Cormack (2008). "Disparity Statistics in Natural Scenes." Journal of Vision **8**: 1-14.
Loomis, J. M., J. a. Da Silva, N. Fujita and S. S. Fukusima (1992). "Visual space perception and visually directed action." Journal of experimental psychology. Human perception and performance **18**: 906-921.

Loomis, J. M., J. a. da Silva, J. W. Philbeck and S. S. Fukusima (1996). "Visual Perception of Location and Distance." Current Directions in Psychological Science **5**: 72-77.

Loomis, J. M. and J. W. Philbeck (1999). "Is the anisotropy of perceived 3-D shape invariant across scale?" Perception & psychophysics **61**: 397-402.

Love, G. D., D. M. Hoffman, P. J. Hands, J. Gao, A. K. Kirby and M. S. Banks (2009). "High-speed switchable lens enables the development of a volumetric stereoscopic display." Opt Express **17**(18): 15716-15725.

Maloney, L. T. and M. S. Landy (1989). "A Statistical Framework for Robust Fusion of Depth Perception." Visual Communications and Image Processing IV **1199**: 1154-1163.

Marr, D. and T. Poggio (1979). "A computational theory of human stereo vision." Proc R Soc Lond B Biol Sci **204**(1156): 301-328.

Müller, J. (1826). Comparative Physiology of the Visual Sense. Cnobloch Leipzig.

Oruç, İ., L. T. Maloney and M. S. Landy (2003). "Weighted linear cue combination with possibly correlated error." Vision Research **43**: 2451-2468.

Palmer, S. E. (1999). Vision science : photons to phenomenology. Cambridge, Mass., MIT Press.

Palmisano, S., B. Gillam and D. Govan (2010). "Stereoscopic perception of real depths at large distances." Journal of vision **10**: 1-16.

Philbeck, J. W. and J. M. Loomis (1997). "Comparison of two indicators of perceived egocentric distance under full-cue and reduced-cue conditions." Journal of experimental psychology. Human perception and performance **23**: 72-85.

Rao, R. P. (1999). "An optimal estimation approach to visual perception and learning." Vision research **39**: 1963-1989.

Robson, J. G. and N. Graham (1981). "Probability summation and regional variation in contrast sensitivity across the visual field." Vision Res **21**(3): 409-418.

Rust, N. C., & Movshon, J. A. (2005). "In praise of artifice." Nature neuroscience, **8**(12): 1647-1650.

Saunders, J. A. (2003). "The effect of texture relief on perception of slant from texture." Perception **32**(2): 211-233.

Shibata, T., J. Kim, D. M. Hoffman and M. S. Banks (2011). "Visual discomfort with stereo displays: Effects of viewing distance and direction of vergence-accommodation conflict." Proc SPIE **7863**: 78630P78631-78630P78639.

Stockman, A., L. T. Sharpe, S. Merbs and J. Nathans (2000). "Spectral sensitivities of human cone visual pigments determined in vivo and in vitro." Methods Enzymol **316**: 626-650.

Westheimer, G. and S. P. McKee (1977). "Spatial configurations for visual hyperacuity." Vision Res **17**(8): 941-947.

Wheatstone, C. (1838). "Contributions to the physiology of vision.--Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision." Philosophical transactions of the Royal Society of London: 371-394.

Yang, Z. and D. Purves (2003). "A statistical explanation of visual space." Nature neuroscience **6**: 632-640.

Zhang, Z. (2000). "A flexible new technique for camera calibration." Pattern Analysis and Machine Intelligence, IEEE Transactions on **22**(11): 1330-1334.

Vita

Brian Clark McCann was born in Portland, Maine. His interest in graphics and artificial intelligence began at Lincoln-Sudbury Regional High School, Sudbury, Massachusetts (graduation 1999). After graduating, he went on to pursue his Bachelor's of Science in Computer Science with focus on computer vision from the University of Rochester, Rochester, NY (graduation 2003). For the next four years, Brian stayed at the University as a programmer in the lab of David C. Knill in the center for visual sciences. In 2007, Brian began his doctoral program in Perception with the Department of Psychology and the Center for Perceptual Systems at the University of Texas at Austin. Following graduation, Brian plans to stay at the University as a Research Associate for the Texas Advanced Computing Center.

Email: brian.c.mccann@utexas.edu

This dissertation was typed by Brian Clark McCann